MeCAS
Maine's Comprehensive
Assessment System

Maine High School Assessment

MeCAS Part II

2013–14 Technical Report

measured
progress

100 Education Way, Dover, NH 03821 (800) 431-8901
WWW.MEASUREDPROGRESS.ORG

TABLE OF CONTENTS

CHAPTER 1 OVERVIEW OF THE MAINE HIGH SCHOOL ASSESSMENT

1.1 PURPOSE OF THE ASSESSMENT SYSTEM

The Maine High School Assessment (MHSA) is designed to measure student progress toward the achievement of the state academic standards contained in Maine's system of *Maine's Learning Results: Parameters for Essential Instruction*. The *Learning Results* content standards are designed to identify the skills and knowledge that all Maine students will need to succeed in the 21st century and are intended to provide them the opportunity to be ready for college, career, and citizenship upon graduation.

Since the spring of 2006, these academic standards have been measured by the SAT. All Maine third-year high school students have been required to participate in the state's SAT Initiative Program, which produces individual measures in mathematics, critical reading, and writing. The SAT is administered annually to Maine students on the first Saturday in May, with the official makeup opportunity administered on the first Saturday in June.

Beginning in 2007–08, a fourth discipline, science, was added to the MHSA compilation as required under No Child Left Behind (NCLB) and is administered in each Maine high school on a school day(s) across a two-week administration window in early April. The same administration protocols and time lines as described above were followed in 2013–14.

As in previous years, all Maine public high schools were designated as SAT test centers. One school opted to send students to a nearby high school/test center.

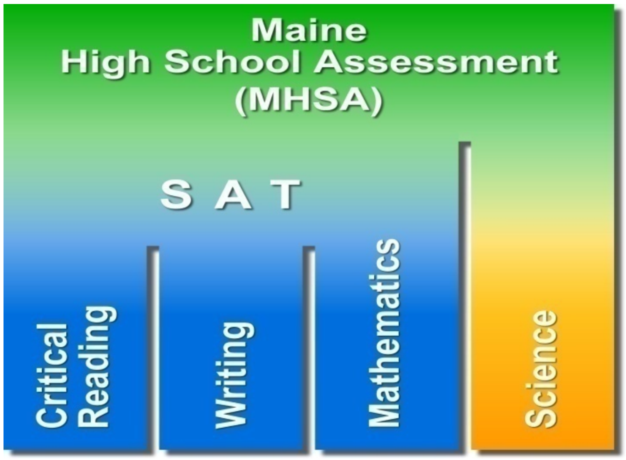
Students who were approved for accommodations received the same accommodations on all components of the MHSA, as explained in Chapter 4. Details about the administration of the science components and the SAT were communicated to schools on an ongoing basis through informational letters, the Maine Department of Education (MDOE) Web site, and webinars. The webinar contained information on all aspects of accommodations, the registration process, SAT test center supervisor training, and science administration.

After the May and June SAT administrations, students testing under standard conditions or with College Board–approved accommodations received official SAT score reports from the College Board. Additionally, *all* students participating in the MHSA received individual score reports based on Maine's achievement levels. The MHSA scores were then used for accountability purposes.

The two components of the 2014 MHSA (SAT and science) comprised a cohesive system with comparable item development, administration, and scoring protocols; similar test material formats; the same accommodations; and a seamless reporting system. Collaboration between the MDOE, the College Board, and

Measured Progress assured that the entire process worked smoothly. This illustration has been used in public presentations to communicate the relationship between the SAT and the complete MHSA program.

Figure 1-1. 2013–14 MHSA: Content Areas



1.2 Purpose of This Report

The purpose of this report is to document the technical aspects of the 2013–14 MHSA, one component of Maine's Comprehensive Assessment System (MeCAS). Other components are the New England Common Assessment Program (NECAP), the Maine Educational Assessment (MEA), and the Personalized Alternate Assessment Portfolio (PAAP), each of which is documented in a separate report.

This report provides information about the technical quality of the MHSA, including a description of the processes used to develop, administer, and score the test and to analyze the test results. It is intended to serve as a guide for replicating and/or improving the procedural and analytical processes to be followed in subsequent years for the MHSA component of Maine's testing program. It was written by staff at the College Board, the SAT contractor, and Measured Progress, the MHSA testing contractor; reviewed by members of the Maine Technical Advisory Committee for Assessment (see Appendix A); and edited by MDOE staff.

While some sections of this technical report may be used by educated laypeople, it is intended for experts in psychometrics and educational research. The report assumes a working knowledge of measurement concepts such as *reliability* and *validity*, and statistical concepts such as *correlation* and *central tendency*. In some chapters, the reader is presumed also to have basic familiarity with advanced topics in measurement and statistics.

CHAPTER 2 TEST DESIGN AND DEVELOPMENT OF THE MHSA: SAT

The MHSA is intended to support good educational practice and is perceived as having an impact on instruction and curriculum. It features two components: The SAT and the MHSA Science test. Details on the content specifications and development of the SAT component are featured in this chapter; Chapter 3 covers content specifications and development of the Science component.

The SAT Committee—composed of teachers, academic administrators, measurement experts, admissions officers, college counselors, and students—provides the College Board with advice on any of the policies, practices, products, and services involving the SAT. In addition, the development of each of the three content areas on the SAT (mathematics, critical reading, and writing) is guided by the work of a test development committee composed of both secondary school and college teachers in that content area. The involvement of these development committees will be identified in the discussion of the test development process below. The current members of these committees can be found at www.collegeboard.org.

2.1 THE MHSA: THE SAT OVERVIEW

Detailed content and statistical specifications for each of the three content areas define the parameters that ensure that each new form is comparable to all other forms of the SAT. That is, the detailed test specifications and statistical procedures ensure that different forms of the same test developed both within each academic year and across years are parallel in content and difficulty. These design features, plus SAT equating procedures, enable comparability of scores from different test administrations. For example, Maine scores from the May 2014 administration of the SAT can be directly compared with scores from the May 2013 administration. The MHSA designates the May and June (makeup only) SAT administration dates for state assessment purposes. Scores from these administrations can also be directly compared. The specifications for both the content and the psychometric characteristics of each test are provided later in this chapter. Examples of each type of question used on the test may be found at www.collegeboard.org.

2.2 UNIVERSAL DESIGN SPECIFICATIONS

The SAT components of the MHSA are developed according to the following six principles of universal design defined by Thompson, Johnstone, and Thurlow (2002):

1. Inclusive assessment population—The MHSA: SAT provides assessment opportunities for all students, regardless of their cognitive abilities, cultural backgrounds, or linguistic backgrounds.

2. Precisely defined constructs—The MHSA: SAT measures the constructs it is intended to measure and does not measure irrelevant material.

3. Accessible, non-biased items—The MHSA: SAT uses appropriate accommodations to "level the playing field" for students with disabilities. These accommodations do not affect the validity of the assessments or the comparability of scores obtained on them.

4. Simple, clear, and intuitive instructions and procedures—The MHSA: SAT instructions are easy to understand regardless of a student's experience, knowledge, language skills, or current concentration level. In addition, test development committees review SAT instructions to ensure that they are appropriate for the test-taking population.

5. Maximum readability and comprehensibility—MHSA: SAT mathematics items are developed with the minimal number of required words and the least amount of grammatical complexity for the task. For the critical reading and writing items, the level of readability and syntax is appropriate for the construct that is being measured by those items. Readability is part of the thorough review by content experts before and after the pretesting of items.

6. Maximum legibility—The text, tables, and figures that appear on the MHSA: SAT are designed to ensure maximum legibility. In the mathematics sections, figures that accompany problems are intended to provide information useful in solving the problems. All figures are drawn to scale unless otherwise indicated.

2.3 SAT Critical Reading Test

The May and June 2014 forms required by the MHSA, like all forms of the SAT critical reading test, met the specifications presented in Table 2-1.

Table 2-1. 2013–14 MHSA: SAT Critical Reading Content Specifications

	Number	*Percentage of Test*
Time allotted	70 minutes	
Sentence completion	19 items	28
Passage-based reading	48 items	72
Total	67 items	100
800-word passages*	2 passages	
650-word passages*	1 passage	
500-word passages*	1 passage	
Paragraph reading	2 passages	
Paired paragraph	1 pair	
Extended reasoning	36–40 items	54–60
Literal comprehension	4–6 items	6–9
Vocabulary in context	4–6 items	6–9

* Note: One of the long passages will actually be a pair of related passages (e.g., instead of an 800-word passage, there will be two related 400-word passages, etc.)

Each new form of the SAT critical reading test will continue to meet the listed specifications.

The passage-based reading content is balanced across four categories: humanities, social studies, natural sciences, and literary fiction. Male and female references are balanced across the test. Representative minority-relevant content is included. Approximately 80% of the passage-based reading content (60% of the total test) measures extended reasoning skills through questions about primary purpose, rhetorical strategies, implication and evaluation, tone and attitude, application and analogy; the balance of the questions are concerned with literal comprehension or vocabulary in context. The three separately timed sections of a typical SAT critical reading test are configured as shown in Table 2-2.

An important constraint in the development of multiple parallel forms of a test is that the distribution of item difficulties be the same across forms. Using the equated delta[1] index, each SAT critical reading test must have questions with the distribution of difficulty indicated in Table 2-3.

Table 2-2. 2013–14 MHSA: SAT Critical Reading Section Configuration

Reading 1 (25 minutes)	*Reading 2 (25 minutes)*	*Reading 3 (20 minutes)*
Items 1–8: Sentence completion items (8)	Items 1–5: Sentence completion items (5)	Items 1–6: Sentence completion items (6)
Items 9–12: Either two paragraph reading passages with two items each OR one paired paragraph with four items (4)	Items 6–9: Either two paragraph reading passages with two items each OR one paired paragraph with four items (4)	Items 7–19: One 800-word passage with 13 items
Items 13–24: One 800-word passage with 12 items	Items 10–24: One 500-word passage and one 650-word passage with a total of 15 items	

Note: The actual number of passage-based reading questions in each section may vary by one or two, but the total number in each critical reading test will always be 48.

Table 2-3. 2013–14 MHSA: SAT Critical Reading Psychometric Specifications

	Item Type Difficulty	
	Sentence Completion	*Passage-based Reading*
Mean equated Delta by item type	10.4–12.4	10.4–12.4
Equated Delta Distribution for the Overall Test		
Mean equated delta (SD)	11.4 (2.4)	

Number and Percentage of Items by Delta Value		
DV	*N*	*(%)*
16	1	(1.5)
15	4	(6.0)
14	6	(9.0)
13	7	(10.4)
12	9	(13.4)
11	12	(17.9)
10	9	(13.4)

continued

[1] Described more fully in Chapter 8, equated delta is a transformation of 1–p, with a mean of 13 and a standard deviation of 4.

DV	*N*	*(%)*
9	7	(10.4)
8	6	(9.0)
7	4	(6.0)
6	2	(3.0)
Total	67	(100)

Note: The equated delta distribution, mean, and standard deviation are provided for the overall reading test, while the equated delta mean is provided for the two item types. It is not necessary to specify the standard deviation of the mean equated delta by item type because the reading test is assembled to meet the overall point by point delta distribution.

2.4 SAT Writing Test

Although Maine does not use writing as an adequate yearly progress (AYP) measure for accountability under NCLB, Maine includes writing in its assessment system. The May and June 2014 forms required by the MHSA, like all forms of the SAT writing test, met the specifications presented in Table 2-4:

Table 2-4. 2013–14 MHSA: SAT Writing Content Specifications

Time Allotted-60 minutes	*Number*	*Percent of MC*[1] *Portion*
Improving sentences (sentence correction)	25 items	51
Identifying sentence errors (usage)	18 items	37
Improving paragraphs (revision in context)	6 items based on a passage[2]	12
Total	49 items	100
Essay	1 essay	

[1] MC = multiple-choice

[2] Passages can range from 150 to 250 words.

Each new form of the SAT writing test will continue to meet the listed specifications.

The essay portion of the test requires students to write an original first draft of an essay in which they develop a point of view on an issue that has been presented through a prompt. The prompt is written to be easily accessible to the general test-taking population, including students for whom English is a second language, and is free of figurative or technical language or specific literary references. The prompt presents an issue that engages students of high school age and allows them to draw on their knowledge and interests to respond. The prompt outlines a range of possible viewpoints within a single issue, and stimulates critical reflection on the issue. Following the prompt is an assignment that focuses the student on the issues addressed in the prompt. The essay is scored by trained readers using the essay scoring guide, displayed as Figure 2-1.

Figure 2-1. 2013–14 MHSA: Essay Scoring Guide

ESSAY SCORING GUIDE

The Scoring Guide expresses the criteria readers use to evaluate and score the student essays. The Guide is structured on a six-point scale. The language of the Scoring Guide provides a consistent and coherent framework for differentiating between score points, without defining specific traits or types of essays that define each score point.

Score of 6

An essay in this category demonstrates **clear and consistent mastery**, although it may have a few minor errors. A typical essay

- effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position
- is well organized and clearly focused, demonstrating coherence and smooth progression of ideas
- exhibits skillful use of language, using a varied, accurate, and apt vocabulary
- demonstrates meaningful variety in sentence structure
- is free of most errors in grammar, usage, and mechanics

Score of 5

An essay in this category demonstrates **reasonably consistent mastery**, although it will have occasional errors or lapses in quality. A typical essay

- effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position
- is well organized and focused, demonstrating coherence and progression of ideas
- exhibits facility in the use of language, using appropriate vocabulary
- demonstrates variety in sentence structure
- is generally free of most errors in grammar, usage, and mechanics

Score of 4

An essay in this category demonstrates **adequate mastery**, although it will have lapses in quality. A typical essay

- develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position
- is generally organized and focused, demonstrating some coherence and progression of ideas
- exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary
- demonstrates some variety in sentence structure
- has some errors in grammar, usage, and mechanics

Score of 3

An essay in this category demonstrates **developing mastery**, and is marked by one or more of the following weaknesses:

- develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position
- is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas
- displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice
- lacks variety or demonstrates problems in sentence structure
- contains an accumulation of errors in grammar, usage, and mechanics

Score of 2

An essay in this category demonstrates **little mastery**, and is flawed by one or more of the following weaknesses:

- develops a point of view on the issue that is vague or seriously limited and demonstrates weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position
- is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas
- displays very little facility in the use of language, using very limited vocabulary or incorrect word choice
- demonstrates frequent problems in sentence structure
- contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured

Score of 1

An essay in this category demonstrates very little or no mastery, and is severely flawed by one or more of the following weaknesses:

- develops no viable point of view on the issue, or provides little or no evidence to support its position
- is disorganized or unfocused, resulting in a disjointed or incoherent essay
- displays fundamental errors in vocabulary
- demonstrates severe flaws in sentence structure
- contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning

Score of 0

Essays not written on the essay assignment will receive a score of zero.

As illustrated in Table 2-5, writing process skills are assessed through both the improving paragraphs item type and through the essay that each student writes.

Table 2-5. 2013–14 MHSA: Alignment Between Writing Process Skills and SAT Writing Questions

Writing Process Skill	*Essay Prompt*	*Improving Paragraphs*
Writing personal narratives	X	
Using literal and figurative language appropriately	X	X
Using sentence variety	X	X
Demonstrating insight and/or creativity in the writing task	X	
Using topic sentences	X	X
Using appropriate voice, tone, and style	X	X
Focusing on a purpose for writing	X	
Writing persuasive and/or argumentative essays	X	
Organizing paragraphs and using appropriate transitions	X	X
Writing effective introductions and conclusions	X	X
Using writing and reading as tools for critical thinking	X	
Developing a logical argument	X	
Writing a unified essay	X	X
Using supporting details and examples	X	
Writing a clear and coherent essay	X	X

The multiple-choice writing questions test a wide range of grammatical, usage, and sentence-structure skills as shown in Table 2-6.

Table 2-6. 2013–14 MHSA: Alignment Between Grammar, Usage, and Sentence-Structure Skills and the Problems Tested by SAT Writing Questions

Grammar, Usage, and Sentence-Structure Skill	*Improving Sentences*	*Identifying Sentence Errors*	*Improving Paragraphs*
Avoiding faulty predication in sentences	X	X	X
Avoiding dangling modifiers	X		
Using comparative modifiers appropriately	X	X	
Using appropriate idiomatic words, phrases, or structures	X	X	X
Avoiding weak, passive constructions	X		
Using connectives appropriately	X	X	X
Avoiding illogical comparisons	X	X	
Subordinating and coordinating ideas in sentences	X	X	X
Avoiding pronoun shift	X	X	X
Combining sentences appropriately			X
Maintaining parallel structure in sentences	X	X	X
Using appropriate verb forms	X	X	X
Avoiding wordiness	X	X	X
Controlling errors in subject-verb agreement	X	X	
Avoiding errors in pronoun agreement, case, and reference	X	X	
Maintaining tense sequences	X	X	X
Making acceptable word choices	X	X	X
Avoiding run-on sentences	X		
Avoiding sentence fragments	X		X
Avoiding comma splices	X		X

The SAT writing test is administered in three separately timed sections as configured in Table 2-7.

Table 2-7. 2013–14 MHSA: SAT Writing Section Configuration

Writing 1 (25 minutes)	*Writing 2 (25 minutes)*	*Writing 3 (10 minutes)*
Essay	Items 1–11: Improving sentences (11) Items 12–29: Identifying sentence errors (18) Items 30–35: Improving paragraphs (6)	Items 1–14: Improving sentences (14)

Multiple-choice items are spread across a variety of content areas, including science, practical affairs, human relations, geography, literature, art, legal, education, business, and history. Female and male references are balanced, and representative minority-relevant content is included.

In order to develop multiple parallel forms of a test, the distribution of item difficulties must be the same across forms. Using the equated delta index, each section of the SAT writing multiple-choice portion of the test must have questions with the distribution of difficulty indicated in Table 2-8.

Table 2-8. 2013–14 MHSA: SAT Writing Psychometric Specifications

Equated Delta Distribution for the Multiple-Choice Portion of the Test		
Mean equated delta (SD)	10.1 (2.5)	
Number and Percentage of Items by Delta Value		
DV	*N*	*(%)*
16	1	(2.0)
15	0	(0.0)
14	2	(4.1)
13	3	(6.1)
12	6	(12.2)
11	7	(14.3)
10	7	(14.3)
9	7	(14.3)
8	6	(12.2)
7	5	(10.2)
6	3	(6.1)
5	2	(4.1)
Total	49	(100)

2.5 MHSA MATHEMATICS TEST: SAT

The MHSA mathematics test consists of the traditional SAT mathematics. The content specifications for the SAT component remain relatively stable from year to year, with only slight differences due to a range of acceptable numbers of items measuring particular content specifications and routine variability as to whether the test form fell on the upper or lower end of the acceptable range. For a small number of content specifications, an item measuring that content may or may not be included on every form.

The May and June 2014 SAT forms required by the MHSA, like all forms of the SAT mathematics test, met the specifications presented in Table 2-9.

Table 2-9. 2013–14 MHSA: SAT Mathematics Content Specifications

Time Allotted: 70 minutes	*Number*	*Percent of Test*
Multiple-choice	44 items	81
Student-produced response	10 items	19
Total	54 items	
Number and Operations	11–13 items	20–24
Algebra and Functions	19–21 items	35–39
Geometry and Measurement	14–16 items	26–30
Data Analysis, Statistics, and Probability	6–7 items	11–13

Each new form of the SAT mathematics test will continue to meet the specifications listed. The four content areas specified in Table 2-9 are further defined in Table 2-10.

Table 2-10. 2013–14 MHSA: SAT Mathematics Content Description

Number and Operations
• Arithmetic word problems (including percent, ratio, and proportion)
• Properties of integers (odd/even, prime numbers, divisibility, and so forth)
• Rational numbers
• Logical reasoning
• Sets (union, intersection, elements)
• Counting techniques
• Sequences and series (including exponential growth)
• Elementary number theory
Algebra and Functions
• Substitution and simplifying algebraic expressions
• Properties of exponents
• Algebraic word problems
• Solutions of linear equations and inequalities
• Systems of equations and inequalities
• Quadratic equations
• Rational and radical equations
• Equations of lines
• Absolute values
• Direct and inverse variation
• Concepts of algebraic functions
• Newly defined symbols based on commonly used operations
Geometry and Measurement
• Area and perimeter of a polygon
• Area and circumference of a circle
• Volume of a box, cube, and cylinder
• Pythagorean Theorem and special properties of isosceles, equilateral, and right triangles
• Properties of parallel and perpendicular lines
• Coordinate geometry
• Geometric visualization
• Slope
• Similarity
• Transformations
Data Analysis, Statistics, and Probability
• Data interpretation
• Descriptive statistics (mean, median, mode)
• Probability

The three separately timed SAT mathematics sections are configured as follows:

Table 2-11. 2013–14 MHSA: SAT Mathematics Section Configuration

Mathematics 1 (25 minutes)	*Mathematics 2 (25 minutes)*	*Mathematics 3 (20 minutes)*
Items 1–20: Multiple-choice (20)	Items 1–8: Multiple-choice (8) Items 9–18: Student-produced response (10)	Items 1–16: Multiple-choice (16)

Calculators are permitted on the SAT mathematics test, and basic geometric reference information is provided at the top of each separately timed section. Additional information on the calculator policy for the SAT is provided in Chapter 4.

In order to develop multiple parallel forms of a test, the distribution of item difficulties must be the same across forms. Using the equated delta index, each SAT mathematics test must have questions with the distribution of difficulty indicated in Table 2-12.

Table 2-12. 2013–14 MHSA: SAT Mathematics Psychometric Specifications

	Item Type Difficulty	
	MC	*SPR*
Mean equated delta (SD)	12.2 (3.2)	13.6–14.2 (3)

Number of Items by Delta value and Item Type			
MC		*SPR*	
18–20	1	18–20	1
17	2		
16	2	16–17	2
15	4		
14	5	14–15	2
13	5		
12	5	12–13	2
11	5		
10	4	10–11	2
9	3		
8	3	8–9	1
7	2		
6	2	<8	0
<6	1		
Total	44	Total	10

MC = multiple-choice; SPR = student-produced response; SD = standard deviation

2.6 DEVELOPMENT

Each new form of the MHSA test is developed through a multistage process that spans many months. The basic steps are similar for each of the three content areas (mathematics, critical reading, and writing), although the details of the process may vary somewhat among these three. Significant variations will be noted here as appropriate. The development process draws on the skills of content experts, psychometricians, and experienced educators in order to repeatedly develop new forms that are parallel, fair to students, and test the reasoning skills important to academic success in college. Experienced educators participate in the development process through the work of multiple committees. The current members of these committees can be found at www.collegeboard.org.

2.7 Item Writing and Review

Test development specialists at Educational Testing Service (ETS) write the test items for the SAT. Some of the items are based on ideas from high school and college faculty and other qualified consultants. Faculty and consultants are selected for their knowledge of curriculum and for their expertise in a field. In general, the staff who work on a particular test are content specialists who have either high school or college teaching experience. In writing items, these people are guided by the content and statistical specifications for the particular portion of the MHSA (mathematics, critical reading, or writing) on which they are working.

Because such a high proportion of the questions on the critical reading test are tied to a reading passage, potential reading passages are first chosen and reviewed for suitability before any passage-based items are written. Each newly written item (or set of items) is classified according to the appropriate category of the specifications. It is reviewed to maximize clarity and to eliminate ambiguity. It is further reviewed for sensitivity to members of gender and racial or ethnic subgroups. Each item is also examined to make sure that it has only a single correct answer. The student-produced-response items in mathematics may have more than one possible answer or more than one way to express the answer (see Chapter 4 for more information on student-produced-response items). During the review process, items may be discarded, accepted, or revised to eliminate ambiguity, improve wording, strengthen the correct answers, and so forth.

2.8 Pretesting the Items

Every item used in an operational form of the MHSA SAT has been pretested; that is, the item has been tried out with an appropriate group of students to make sure that it is not ambiguous or confusing and to determine the difficulty level and the degree to which it differentiates more or less able students. The pretest responses are also analyzed to determine whether students of different racial/ethnic or gender groups respond to the question differently. MHSA SAT item writing and review are ongoing activities throughout the year.

The multiple-choice items of the SAT (mathematics, critical reading, and writing), as well as the student-produced-response mathematics items, are pretested on a sample of actual SAT test takers. There are 10 separately timed sections in each SAT: three for the writing test, three for the critical reading test, and three for the mathematics test; the remaining section does not count toward the student’s score and is used either for pretesting, for providing calibration information for the equating of test scores, or for research. Pretests, each configured like one of the operational sections, are assembled from questions that have received a number of content, fairness, and editorial reviews prior to pretesting. Each pretest is administered as the unscored section of some fraction of all SATs administered on a particular date; that is, every nth test book will have a particular pretest or equating test in that unscored section. This pattern of administration provides item information on a large random sample of SAT test takers. Consequently, this item information provides an extremely accurate estimate of how the item will function when administered as part of a future SAT.

Each SAT writing essay prompt is reviewed by SAT staff at both the College Board and ETS. After all concerns raised during the review process are resolved, the essay prompt is pretested in a special administration in high school English classrooms. For each group of pretests, a diverse sample of schools is invited to participate by having students respond to a particular prompt during their English class. A sample of at least 300 responses to each essay prompt is obtained in order to determine whether the question is accessible to students and to provide exemplars of various levels of writing competence for use in the scoring process, described in Chapter 6.

2.9 Analysis of Pretest Information for the MHSA: SAT

Data collected from multiple-choice and student-produced-response pretests are analyzed to provide important information about the appropriateness of items for use in operational forms of the SAT. Three statistical indices are computed: **equated delta** as an index of item difficulty within the SAT population, ***r*-biserial** as an index of whether the item discriminates between more and less able students, and Mantel-Haenszel **DIF** (differential item functioning) as an index of the relationship between group membership and the likelihood of answering the question correctly. These item statistics are used to judge whether a given question is suitable for inclusion in the pool of items from which operational forms are assembled. The item statistics may also reveal problems with the conceptualization or wording of a question. Some of these items are revised and re-pretested. Others are discarded. SAT items are analyzed by ETS using data from the national administration of the test form. The statistical indices employed in analyzing and screening the MHSA SAT components follow.

2.10 Item Difficulty

The difficulty of an item is a function of the percentage of test takers who answer it correctly (i.e., *p*-value). An item’s difficulty should be appropriate for the population taking the test. When an item is too easy, virtually all test takers answer it correctly; thus, extremely easy items contribute very little information to the total test score. Similarly, inappropriately difficult items are not very useful in a test. Because items within a test are highly inter-correlated, it is best to select items with a moderate spread of difficulty around a mean *p*-value of 0.5 (or 50% correct). The required distribution of item difficulty for each part of the SAT is defined in the psychometric specifications found in Tables 2-3, 2-8, and 2-12.

Typically, *p*-values are converted to a standard scale that avoids negative values and decimals (Anastasi, 1976). The measure of difficulty used with the SAT is the delta index (”). This index is based on the percentage of test takers answering a given item correctly, where 1 minus the *p*-values are converted to z-scores and transformed to a scale with a mean of 13 and a standard deviation of 4. The delta scale is inversely related to the *p*-scale; thus, the more difficult the item, the greater the delta value and the smaller the *p*-value.

Because the samples to which specific pretest items are administered in a non-scored section may, to some degree, differ in ability level from the 1990 standard reference group used for the SAT, it is necessary to convert the raw delta values to equated delta values. To make this conversion, data from items in the scored sections is used, since the equated delta values for all of those items are known. The raw delta values for the common items based on the current sample are then plotted against the known equated deltas from the previous equating. The resulting linear relationship between the pairs of raw and equated deltas is used to compute an equated delta for the new pretest item. An equated delta value is computed for each pretest item and is based on the standard reference population, permitting comparisons of items among samples (Thurstone, 1947).

Each form of the SAT is built to a well-defined distribution of item difficulty. While formula scored items include a correction for guessing, the delta scale (based on percentage correct) does not adjust for incorrect responses. As a result, the proportion-correct delta scale provides an estimate of difficulty that is slightly lower than it would be if the formula scoring were taken into account. This is not a problem for the reading test and the multiple-choice portion of the writing test. All items in these sections are formula scored with the same amount subtracted (¼ of a point) for an incorrect response (i.e., the *k*-factor is 0.25), and the statistical specifications have been designed to reflect this known difference. The mathematics test, however, contains both formula-scored multiple-choice items (with a *k*-factor of 0.25) and student-produced-response items that do not penalize incorrect responses. For this reason, as shown in Table 2-12, psychometric specifications for the SAT mathematics test provide separate delta distributions for multiple-choice and student-produced-response items. For more detail on how statistical specifications were set for the SAT, see Lawrence and Schmitt (1994).

2.11 ITEM DISCRIMINATION AND ITEM TEST RELATIONSHIP

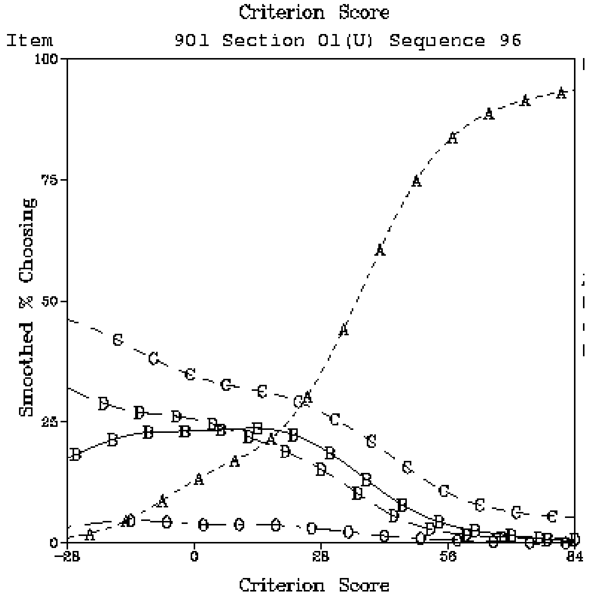
Although difficulty level is one important criterion in selecting items, item discrimination is essential to be able to distinguish among test takers at different levels of ability. The *r*-biserial correlation coefficient between the item and the total test score is most often used to assess the item’s utility in discriminating among test takers of differing ability levels and the homogeneity of test items (or extent that a student’s performance on an item relates to his/her total test score). The biserial correlation ranges from 1 to -1. The more positive the correlation, the more the item distinguishes test takers with high total scores from those with low scores. A negative biserial correlation indicates that the item is measuring something different from the rest of the test; test takers with high scores are more likely to answer that item incorrectly than those with low scores.

Correlations that are near 0 indicate that high scorers and low scorers have the same chance of correctly answering the item. Because of these results, the MHSA does not include items with low or negative biserial correlations.

Biserial correlations also provide an indication of the homogeneity of test items. If the correlation is very close to 1, all of the information provided by the item is redundant with that provided by the other test items. Items with moderate biserial correlations distinguish among ability levels, yet also supply unique information. Therefore, most items included on SAT operational forms fall within a biserial range of 0.30 to 0.80.

In determining whether to select, omit, or edit and refine an item based on results from pretests, test developers also consider the number and percentage of test takers who respond to the correct option and to each incorrect option (with all items on the SAT except student-produced responses and the essay). At each score level, the percentage of test takers selecting each option is plotted. For a correct option, it is expected that the percentage of students selecting the option will increase as the test score increases. Figure 2-2 displays an item with this increasing pattern. If the correct option does not display this pattern, the item is carefully reviewed. Similarly, if an incorrect option has this typical increasing pattern, then that option is closely evaluated. As a result of the evaluation, the item may be revised and then re-pretested, or it may be discarded entirely.

Figure 2-2. 2013–14 MHSA: SAT Typical Discrimination Pattern Among Multiple-Choice Response Options, Where Option A Is the Key



2.12 DIFFERENTIAL ITEM FUNCTIONING

Analyses of differential item functioning (DIF) are conducted to identify items that may function differently for members of different groups. DIF analyses compare the performance of two groups of test takers (e.g., males versus females, Asian American test takers versus White test takers) who have been matched on their reading, writing, or mathematical proficiency (SAT mathematics, critical reading, or writing total score[2]) on each item. The underlying assumption in conducting such analyses is that all test takers demonstrating the same level of proficiency in the content area should have similar chances of answering each item correctly regardless of gender, race, or ethnicity. DIF occurs when individuals with similar scores on the SAT critical reading, SAT writing (multiple-choice), or SAT mathematics tests differ notably in their performance on a specific test item (Crone and Schmitt, 1991). The presence of DIF indicates that an item functions differently for one subgroup than for another subgroup of the same proficiency. While the theoretical framework for explaining DIF is not yet well established, the assumption is that items exhibiting high levels of DIF may be measuring factors irrelevant to a test (such as culture) or more than one dimension for which the two groups have different strengths. For example, DIF may result from a mathematical word problem because the question measures language proficiency in addition to mathematical reasoning. One group of test takers may well be stronger in language proficiency. An item like this would be reviewed by one or more experts who have not been involved with the item and who are trained with respect to the construct being tested and item sensitivity. The experts would determine whether the amount of language proficiency required by the item is irrelevant to the dimension of interest, that is, mathematical reasoning.

DIF analyses begin by examining any differences in the performance on each individual item of two comparable groups, referred to as the reference group and the focal group. Typically, DIF analyses for the SAT compare groups based on gender (where males are the reference group and females are the focal group) or ethnicity/race (where White test takers are the reference group and African American, Hispanic, Asian American, or Native American test takers are the focal group). Occasionally DIF analyses are conducted with other groups (e.g., students with disabilities and those without disabilities; students for whom English is a second language [ESL] and non-ESL students). Items with extreme values of DIF—those items favoring one group over another for examinees of the same level of proficiency—undergo further review to determine whether some aspect of what the item is measuring is particularly related to subgroup membership and irrelevant to the dimension being measured. When an item is identified as exhibiting such characteristics, it is either revised and re-pretested or eliminated. The final form of a test rarely includes an item that exhibits sizable DIF. All items with DIF, however, have been reviewed by experts and have been determined to be appropriate for administration.

[2] Groups of test takers are matched on some criterion that reflects the underlying dimension or construct of interest (e.g., critical reading, mathematical reasoning). Typically this “matching criterion” is the total score on the relevant part of the SAT. However, the criterion may vary with the intent of the study. For example, in examining DIF associated with student-produced–response items on the SAT mathematics test, Lawrence, Lyu, and Feigenbaum (1995) used the raw score on 25 quantitative comparison items (an external matching criterion because it did not include the student-produced–response items under study) and the total raw score for SAT mathematics (an internal matching criterion because it included the student-produced–response items under study).

The Mantel-Haenszel (1959) procedure (MH), adapted by Holland and Thayer (1988), is used for DIF analyses with the SAT.[3] This procedure computes a ratio for the conditional probability of successful reference group performance on an item over the conditional probability of successful focal group performance on the item for each score level on the test. Thus, comparisons are made of test takers with equivalent scores (e.g., equivalent proficiency in mathematical reasoning) at each point on the test. Statistically optimal weights are then assigned to each ratio, and they are averaged across all score points. The MH statistic is transformed to the delta (”) scale described previously, and the resulting statistic is referred to as the Mantel-Haenszel delta DIF (MH D-DIF).

The MH D-DIF statistic ranges from negative infinity to infinity, with a value of 0 indicating no DIF. Both the magnitude of the MH D-DIF and a significance test are used to evaluate the presence or absence of DIF. For the SAT, MH D-DIF values are considered

- negligible if they are between 1.0 and -1.0 or are not statistically different from 0 at the 0.05 significance level;
- moderate if they fall between 1.0 and 1.5 or -1.0 and -1.5, or if they are greater than 1.5 or -1.5 and not statistically different from the absolute value of 1.0 at the 0.05 significance level; and
- sizable if they exceed 1.5 or -1.5 and are statistically different from the absolute value of 1.0 at the 0.05 significance level.

Items exhibiting sizable DIF are not included when a test is assembled. Items exhibiting moderate DIF are usually not selected for a final form unless items with negligible DIF are insufficient to meet particular specifications. The average DIF for each group comparison is constrained to be approximately 0 across all test items in a form when an internal matching criterion (e.g., total test score) is used.

2.13 Evaluating Essay Pretests

As indicated previously, essay pretests are administered in classrooms scattered throughout the country. The responses collected from students are read by a group of experienced teachers, including members of the SAT Writing Committee, to determine whether a particular prompt is readily understood by high school students and elicits responses that reflect differing degrees of writing skill. In other words, does the prompt lead to responses that can be scored reliably and that provide differentiation among better and poorer writers? The members of the pretest reading group individually read and score a substantial number of the responses using the essay scoring guide, displayed as Figure 2-1. As a group, they discuss each prompt and decide whether it should be used, revised, or discarded. From the student responses collected during the pretesting, exemplars are chosen for each point on the holistic scoring scale. These serve as anchor papers for

[3] For a complete description of the DIF procedures used by the SAT, see Dorans and Holland (1993).

training and monitoring the experienced high school and college teachers who serve as readers when the essay prompt is administered operationally. The scoring process is described in Chapter 6.

2.14 Assembling the SAT Portion of the MHSA

The ongoing process of writing, reviewing, and pretesting items results in a large pool of acceptable test questions that are ready to be used in future operational forms of the SAT. Each item is classified according to the content and skill specification(s) of the particular test (mathematics, critical reading, and writing) and by the statistical indices generated by the pretest administration. For each of the three parts of the SAT, each item is stored electronically with its associated classifications and statistics. This electronic system can be used to inventory the item pool to identify, for example, particular areas of the specifications where there are insufficient items. Such information can, in turn, guide item-writing assignments.

The electronic system also assists ETS test developers by assembling a draft test that meets the content and psychometric specifications. The test developer then refines the draft test, making sure, for example, that there is a balance of references to women and men, that a particular concept is included that can occur in a variety of contexts (e.g., absolute value), or that one question does not inadvertently provide the answer or a clue to the answer of another question. The test developer then reviews the entire draft test for unintended patterns, e.g., a key run of five Bs. Because each question needs to provide a combination of content and psychometric characteristics, substituting one question for another may lead to the need for a number of other changes in the draft test in order to meet the overall test specifications. After the test developer has completed the draft test, other SAT staff review it. These reviewers consider the same elements as the test assembler, but specifically focus on whether the draft test fully meets both the content and the psychometric specifications for the test, and whether there is an appropriate balance of gender references or subject contexts for reading passages or mathematics problems. There is, again, a review of the test with regard to whether it portrays members of gender or racial/ethnic groups in a sensitive manner and avoids stereotypes. Individual items are reviewed to ensure clarity and lack of ambiguity, and the test as a whole is reviewed to make sure that it is comparable overall to other forms of the SAT. After the resolution of these reviews, the draft test is ready to be reviewed by the SAT test development committees.

2.15 Reviewing the MHSA: SAT Component

Each draft test is reviewed independently by a substantial number of specialists. Members of the test development committees for each area of the test (mathematics, critical reading, and writing) review and discuss each new form of the test. These reviews are performed both by mail and at the site of the committee meeting. The reviews by mail provide time for consideration and reflection on each question and the test as a whole, plus an opportunity for a reviewer to check a reference or to make sure that no wrong answer on a multiple-choice question can be successfully defended as correct. The onsite reviews provide the opportunity for a reviewer to experience the test in much the same fashion as a student, i.e., with time constraints and a sense of pressure. The concerns identified during the review are discussed with the committee and with the staff of the SAT Program, College Board Test Development, and the MDOE. Each concern must be resolved before the test moves into production and printing for its scheduled administration.

2.16 Test Production for the SAT Component

The production of test booklets for any particular administration of the SAT is very complex. Within an administration, multiple forms of the SAT are produced for use in different settings, e.g., the Sunday test centers, the international test centers, Saturday Eastern U.S. centers, Saturday Western U.S. centers. For any given form, multiple variations are created for security reasons and to accommodate the pretest/equating sections. Preparing print-ready copy for each of these distinct test booklets takes several months. Each distinct booklet must be carefully proofread to ensure that it has the correct sections in the correct sequence, and that no typographical errors have been introduced in the composition process.

The actual printing of SAT test books and answer sheets is performed at one of the few printers equipped to protect the security of the tests, to handle the collation of test form variants, and to package and ship the test books and answer sheets to the test centers. The actual administration of the SAT is described in Chapter 4.

2.17 After the SAT Administration

A number of further checks are made after the administration of the SAT and also after the reporting of student scores. A preliminary item analysis of the multiple-choice and student-produced–response questions is done on a sample of the students taking the SAT. The results are used to make sure that each question behaved as expected in terms of the level of difficulty and its ability to differentiate between more and less able students. Items are again analyzed for DIF among subgroups of the population. All reports from test centers of student complaints of ambiguity or incorrectness are reviewed. If the complaint is valid, appropriate action (e.g., dropping the item from scoring) is taken.

After the preliminary analyses and the work of equating the current form(s) to baseline forms have been completed and the essays have been graded, individual tests are scored and reports are issued to the students, their schools, and the designated colleges.

2.18 Public Access to the SAT

A number of forms of the SAT are made public each year. This enables teachers, counselors, admissions officers, students, and parents to be aware of what is tested by the SAT. Such widely available information may be used by teachers in planning curricula, by college faculty in judging how the SAT corresponds to their expectations of students, or by students to verify the accuracy of their scores and in preparing to do their best on the SAT.

Annually, the forms used in four SAT administrations are available through the SAT Question and Answer Service (QAS). This service gives a student a chance to review a copy of the SAT she or he took, a record of the student's answers, the correct answers, and scoring instructions. QAS also includes information about the types of questions and level of difficulty of each question. It does not include a copy of the student's essay, although that can be viewed as part of the online score report or requested in writing. The May SAT form used as part of the 2013–14 MHSA is one of the four released forms and will be available for Maine educators at no cost to inform teaching and learning of the NECAP Grade Level Expectations used in Maine. A link to the released form administered in Maine is also embedded in the MHSA online reporting tool for use by school administrators and classroom teachers.

Some published SAT forms are used as practice tests, either in a print publication or online at http://sat.collegeboard.org. The Web site version of the practice test provides explanations or annotations for each question. Other published SAT forms contribute to the practice questions and explanations that are provided on the Web site. Yet other forms appear in The *Official SAT Online Course*™ and *The Official SAT Study Guide*™, *Second Edition* both of which include extensive explanations of questions. Copies of *The Official SAT Study Guide* have been provided to all Maine high schools, and *The Official SAT Online Course* is provided at no cost on a year-round basis to all students (grades 9–12), as well as all high school teachers and administrators. In addition to giving explanations for all of the questions on the publicly available forms, SAT program staff also prepare explanations for each Official SAT Question of the Day™ that appears on the Web site.

2.19 ALIGNMENT OF THE SAT TO THE NECAP STANDARDS

In 2009, Maine joined the NECAP consortium and the NECAP Grade Level Expectations were adopted into law. Alignment studies were conducted in August 2009 to compare the content of the SAT with the Grade Level Expectations. The studies revealed that the alignment between the reading and mathematics Grade Level Expectations and those of the SAT critical reading test and mathematics test fully satisfied the criteria of the Webb alignment model. The alignment studies can be accessed at http://www.maine.gov/education/mhsa/supportingdocs.html. The SAT test specifications do not change from one test administration to the next. Due to this stability there is no plan to continue conducting alignment studies on an annual basis so an alignment study was not performed in 2013–14. For historical reference, the alignment protocols used in each year of Maine's SAT Initiative are extensively documented in the *MeCAS Technical Manuals* from 2005–2006 through the present.

2.19.1 Design of SAT Critical Reading

The 2012 NECAP Grade Level Expectations, covered by the SAT critical reading section, include the following:

1. Vocabulary Strategies and Breadth of Vocabulary
2. Initial Understanding of Literary Texts
3. Analysis and Interpretation of Literary Texts/Citing Evidence
4. Analysis and Interpretation of Literary Texts – Author's Craft/Citing Evidence
5. Initial Understanding of Informational Text
6. Analysis and Interpretation of Informational Text/Citing Evidence

The number of items covering each performance indicator section of the reading standard is indicated in Table 2-13.

Table 2-13. 2013–14 MHSA: Number of Items on the SAT Coded to NECAP Grade Level Expectations for Reading

Reading Grade Level Expectation	*SAT Critical Reading (Grade 11)*
Vocabulary Strategies and Breadth of Vocabulary	36
Initial Understanding of Literary Texts	21
Analysis and Interpretation of Literary Texts/Citing Evidence	29
Analysis and Interpretation of Literary Texts – Author's Craft/Citing Evidence	17
Initial Understanding of Informational Text	27
Analysis and Interpretation of Informational Text/Citing Evidence	23

2.19.2 Design of SAT Mathematics

This section addresses only the SAT component of the MHSA Mathematics assessment. The 2012 NECAP Grade Level Expectations for Mathematics covered by the SAT mathematics section include the following:

1. Numbers and Operations
2. Geometry and Measurement
3. Functions and Algebra
4. Data, Statistics, and Probability

Table 2-14 displays the number of SAT items measuring each NECAP Grade Level Expectation.

Table 2-14. 2013–14 MHSA: Number of Items on the SAT Coded to the NECAP Grade Level Expectations for Mathematics

Mathematics Grade Level Expectation	*SAT Mathematics (Grade 11)*
Numbers and Operations	28
Geometry and Measurement	16
Functions and Algebra	39
Data, Statistics, and Probability	7

CHAPTER 3 THE MHSA SCIENCE COMPONENT: DESIGN AND DEVELOPMENT

3.1 TEST SPECIFICATIONS

3.1.1 Criterion-Referenced Test

The MHSA contains a criterion-referenced science test. Items on the science test are developed specifically for Maine and are directly linked to Maine's science content standards. These content standards are the basis for the reporting categories and are used to help guide the development of test items.

3.1.2 Item Types

Maine educators and students are familiar with the types of items used in the assessment program. The types of items and their functions are described below:

- **Multiple-choice** items are used to provide breadth of coverage within a content area. Because they require no more than a minute for most students to answer, multiple-choice items make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills.
- **Constructed-response** items typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—to construct satisfactory responses. Constructed-response items take most students approximately 5–10 minutes to complete.

Note that the use of released MHSA items to prepare students to respond to multiple-choice and constructed-response items is appropriate and encouraged.

3.1.3 Description of Test Design

The science test is structured using both common and field-test items. Common items are taken by all students. Student scores are based only on common items. Field-test items are divided among the forms of the test for each grade level. Each student takes only one form of the test and therefore answers a fraction of the field-test items. Field-test items are not identifiable to test takers and have a negligible impact on testing time. Because all students participate in the field test, it provides the minimum sample size (750–1,500 students per item) needed to produce reliable data that can be used to inform item selection for future tests.

3.2 SCIENCE TEST SPECIFICATIONS

3.2.1 Standards

The 2013–14 MHSA science test items are aligned to the content standards D: The Physical Setting and E: The Living Environment, which are Maine's accountability standards and are described in the Science and Technology section of *Maine's Learning Results: Parameters for Essential Instruction*. No other science content standards are subject to statewide assessment. Content specialists use the content standards, performance indicators, and descriptors to help guide the development of test questions, which may address one or more of the performance indicators listed below.

D: The Physical Setting

D1: Universe and Solar System—Students explain the physical formation and changing nature of our universe and solar system, and how our past and present knowledge of the universe and solar system developed.

D2: Earth—Students describe and analyze the biological, physical, energy, and human influences that shape and alter Earth systems.

D3: Matter and Energy—Students describe the structure, behavior, and interaction of matter at the atomic level and the relationship between matter and energy.

D4: Force and Motion—Students understand that the laws of force and motion are the same across the universe.

E: The Living Environment

E1: Biodiversity—Students describe and analyze the evidence for relatedness among and within diverse populations of organisms and the importance of biodiversity.

E2: Ecosystem—Students describe and analyze the interactions, cycles, and factors that affect short-term and long-term ecosystem stability and change.

E3: Cells—Students describe structure and function of cells at the intracellular and molecular levels, including differentiation to form systems, interactions between cells and their environment, and the impact of cellular processes and changes on individuals.

E4: Heredity and Reproduction—Students examine the role of DNA in transferring traits from generation to generation, in differentiating cells, and in evolving new species.

E5: Evolution—Students describe the interactions between and among species, populations, and environments that lead to natural selections and evolution.

3.2.2 Item Types

The science test includes multiple-choice and constructed-response items. Each multiple-choice item requires students to select the correct response from four choices. Each type of item is worth a specific number of points in the student's total science score, as shown in Table 3-1. To prevent being rewarded for guessing, students receive a score of -1/3 on multiple-choice items they answer incorrectly.

Table 3-1. 2013–14 MHSA: Science Item Types

Item Type	*Possible Score Points*
MC	-1/3, 0, or 1
CR	0, 1, 2, 3, or 4

MC = multiple-choice;
CR = constructed-response

Consistent with the annual release policy, 50% of the items were released from the 2013–14 MHSA science test. A practice test composed of released science items is available on the MDOE Web site: http://www.maine.gov/doe/mhsa/resources/released.html. Schools are encouraged to incorporate the use of the released items in their instructional activities so students will be familiar with them.

3.2.3 Test Design

Table 3-2 summarizes the numbers and types of items that were used to compute student scores on the 2013–14 MHSA science test. Additionally, each test form had eight multiple-choice field-test items and one constructed-response field-test item that did not affect student scores.

Table 3-2. 2013–14 MHSA: Science Items

Session 1	*Session 2*	*TOTAL*	
		MC	*CR*
16 MC, 2 CR	24 MC, 2 CR	40	4

MC = multiple-choice;
CR = constructed-response

3.2.4 Blueprints

Table 3-3 shows the distribution of points across the science standards. For MHSA, D1–D2 contained one constructed-response item, D3–D4 contained two constructed-response items, and E1–E5 contained two constructed-response items.

Table 3-3. 2013–14 MHSA: Science Distribution of Score Points

Science Standards	
D1–D2 (Earth & Space)	12
D3–D4 (Physical)	22
E1–E5 (Life)	22
Total score points	56

3.2.5 Depth of Knowledge

Each item on the MHSA science test is assigned a depth-of-knowledge (DOK) level. The DOK level reflects the complexity of mental processing students use to answer an item. DOK is not synonymous with difficulty. Each of the four DOK levels is described below.

- **Level 1 (Recall):** This level requires the recall of information such as a fact, definition, term, or simple procedure. These items require students only to demonstrate a rote response, use a well-known formula, or follow a set procedure.
- **Level 2 (Skill/Concept):** This level requires mental processing beyond that of recalling or reproducing a response. These items require students to make some decisions about how to approach the item.
- **Level 3 (Strategic Thinking):** This level requires reasoning, planning, and using evidence. These items require students to handle more complexity and abstraction than items at the previous two levels.
- **Level 4 (Extended Thinking):** This level requires planning, investigating, and complex reasoning over an extended period of time. Students are required to make several connections within and across content areas. This level may require students to design and conduct experiments. Due to the nature of this level, there are no level 4 items on the MHSA.

It is important that the MHSA measures a range of depths of knowledge. Table 3-4 shows the distribution of points across the DOK levels used on the MHSA.

Table 3-4. 2013–14 MHSA: Science Distribution of Score Points Across Depth of Knowledge (DOK)

DOK Level	*Points*
1	13
2	27
3	16
Total	56

3.2.6 Use of Calculators and Reference Sheets

Calculators are not used or needed when taking the science tests. There are no science reference sheets.

3.3 TEST DEVELOPMENT PROCESS

3.3.1 Item Development

Items used on the science test are developed and customized specifically for use on the MHSA and are consistent with Maine content standards and performance indicators. A Measured Progress test developer works with the Maine state science specialist and Maine educators to verify the alignment of items to the appropriate Maine content standards.

The development process combines the expertise of the Measured Progress test developer, the Maine state science specialist, and committees of Maine educators to help ensure that items meet the needs of the MHSA program. All items used on the common portions of the MHSA were reviewed by a committee of Maine content experts, a committee of Maine bias experts, and three external content experts.

3.3.2 Item Reviews at Measured Progress

The test developers at Measured Progress reviewed newly developed items for

- alignment to the intended content standard;
- item integrity, including science content and structure, format, clarity, possible ambiguity, and single correct answer;
- appropriateness and quality of graphic;
- appropriateness of scoring guide descriptions and distinctions;
- completeness of associated item documentation (e.g., scoring guide, content codes, key, grade level, DOK, and contract identified);
- appropriateness for the designated grade level.

3.3.3 Item Reviews at State Level

A committee of Maine classroom teachers from across the state reviewed the items before field-testing. Teacher participants are selected based on their content-area expertise and grade-level familiarity. The purpose of the review is to evaluate new items for the embedded field test and determine their suitability for the assessment by answering the following four questions:

- Does the item align with the assigned content standard and performance indicator?
- Is the science content accurate?
- Is the science content grade-level appropriate?
- Does the item provide maximum accessibility for all students?

3.3.4 Bias and Sensitivity Review

Bias and sensitivity review is an essential component of the development process. During the bias review process, items were reviewed by a committee of Maine educators who represented various student subgroups, including students with disabilities. Items were examined for content that might cause the test to be inaccessible for these students or that might in general offend or dismay students, teachers, or parents. Being aware of these considerations in the development of assessment items and materials can avoid many unduly controversial issues, and unfounded concerns can be allayed before the test forms are produced.

3.3.5 External Expert Review

The test items were classified into three groups based on science content. Three science experts (one in earth/space science, one in life science, one in physical science) reviewed the group of items corresponding to their area of expertise. The expert reviewers primarily evaluated each item for correct science content. For the multiple-choice items, the experts also indicated whether the keyed answer was correct and whether it was the only correct answer among the options given. The DOE state science specialist and Measured Progress test developers reviewed the experts' evaluations and made appropriate adjustments to the items as necessary.

3.3.6 Reviewing and Refining

Recommended changes from the Item Review and Bias and Sensitivity meetings, as well as the comments from the three external science experts, were reviewed and considered by the Maine state science specialist. Measured Progress test developers made the edits that were approved by the Maine state science specialist.

3.3.7 Item Editing

Measured Progress editors reviewed and edited the items to ensure adherence to sound testing principles and to style guidelines in *The Chicago Manual of Style*, 16th edition. These principles include the stipulations that items

- demonstrate correct grammar, punctuation, usage, and spelling;
- are written in a clear, concise style;
- contain unambiguous explanations that tell students what is required to attain a maximum score;
- are written at a reading level that allows students to demonstrate their knowledge of the subject matter being tested regardless of reading ability;
- exhibit high technical quality regarding psychometric characteristics;
- have appropriate answer options or score-point descriptors; and
- are free of potentially insensitive content.

3.3.8 Item Selection and Operational Test Assembly

Measured Progress test developers met with the Maine state science specialist to select the common items. In preparation for the meeting, the test developers and psychometricians at Measured Progress considered the following in selecting sets of items to propose for the common test:

- **Content coverage/match to test design and blueprints.** The test design and blueprints stipulate a specific number of multiple-choice and constructed-response items. Item selection for the embedded field test was based on the number of items in the existing pool of items that are eligible for the common test.
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously field-tested items were used to ensure quality psychometric characteristics, as well as similar levels of difficulty and complexity from year to year.
- **"Cueing" items.** Items were reviewed for any information that might "cue" or provide information that would help to answer another item.

At the meeting, the Maine state science specialist reviewed the proposed sets of items and made the final selection of items for the common test.

The test developers then sorted and laid out the items into test forms. During assembly of the test forms, the following criteria were considered:

- **Key patterns.** The sequence of keys (correct answers) was reviewed to ensure that their order appeared random.
- **Option balance.** Items were balanced across forms so that each form contained a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- **Page fit.** Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Relationships among forms.** Although field-test items differ from form to form, these items must take up the same number of pages in all forms so that sessions begin on the same page in every form. Therefore, the number of pages needed for the longest form often determines the layout of each form.
- **Visual appeal.** The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of "white space," the density of the test, and the number of graphics.

3.3.9 Operational Test Draft Review

After the forms were laid out as they would appear in the final test booklets, the forms were again thoroughly reviewed by Measured Progress editors to ensure that the items appeared exactly as intended. Any changes made during test construction were reviewed and approved by the test developer. The Maine state science specialist then read the forms for final approval.

3.3.10 Alternative Presentations

The common test for each grade was translated into Braille by the American Printing House for the Blind, a subcontractor that specializes in test materials for blind and visually impaired students. In addition, Form 1 for each grade was adapted into a large-print version.

CHAPTER 4 TEST ADMINISTRATION: SAT

At each administration of the MHSA, great care is taken to ensure that the SAT is administered to all Maine students in a fair, equitable, and standardized manner. The goal of this detailed process is to ensure that all students take the test under a uniform set of conditions so that the results are trustworthy and can be used with confidence in accountability reporting, counseling students, and making admissions and placement decisions. No one is to suffer a disadvantage or gain an advantage of any kind because of race, ethnicity, religion, gender, or disability.

The SAT component of the MHSA was offered to all Maine juniors or third-year students on May 3 and June 7, 2014. There were 12,772 students who registered for and tested on the May date and 157 for the June makeup date. Of those 12,929 students, 12,109 (93.6%) registered under standard conditions, 1,268 (9.8%) registered with College Board–approved accommodations, and 101 (0.7%) preregistered with Maine Purposes Only (MPO) accommodations. On the day of testing, 94 (0.7%) students tested with MPO accommodations and 1,075 (8.3%) students tested with College Board-approved accommodations. Some students moved to MPO accommodations even though they had been approved for accommodations by the College Board because either they were not approved by the College Board for all of the accommodations they requested or they were absent from the May testing and chose to test during the MPO window the following week. The resulting scores were not reportable for college admissions purposes.

4.1 PREPARATION

To promote its goal, the MDOE, in conjunction with the College Board, provides all students planning to take the SAT with extensive practice material in both online and print formats. These range from detailed descriptions of the test, to full-length sample tests, to discussions of approaches to testing, to last-minute tips (e.g., bring a snack) to help each student on the actual test day. The preparatory material may be viewed at www.maine.gov/education/mhsa/studentrp.html and sat.collegeboard.org/practice.

4.2 SUPERVISION

Each Maine public high school where the SAT is administered is supervised by an experienced educator trained by the College Board and provided with detailed instructions and scripts for administering the SAT. The supervisor is responsible for all aspects of the test administration, including hiring staff who meet College Board qualifications, planning the use of the facility, and ensuring the security of test materials from their arrival until their return. The test center staff reflects the diversity of the students being tested and were expected to act in a fair, courteous, nondiscriminatory, and professional manner.

The primary task of all test center staff is to provide an equitable, valid, and standardized test administration. The supervisor is assisted by associate (or room) supervisors and proctors. The associate supervisor checks student identification, reads the test administration script verbatim, and manages all other aspects of the administration in his or her assigned room. In large rooms, the associate supervisor is joined by one or more proctors; the ratio of proctors to students is one additional proctor to every 35–50 students. During the course of the administration, the staff in each room distributes and collects test materials, tells students when to begin and end each test section, walk around the room to guard against misconduct, ensure that each student works on the appropriate section of the test and uses appropriate pencils for marking the answer sheet, and makes sure that no test material is taken from the room.

In addition to standard testing rooms, most test centers have a separate room for students receiving College Board–approved accommodations and/or for students receiving MPO accommodations, which are described later in this chapter. Finally, students whose disabilities cannot be accommodated at the test center (e.g., 100% extended time) are tested (or completed testing) in school the following week using an alternate but comparable form of the SAT.

4.3 Physical Setting

In order that testing takes place in a familiar environment conducive to each student doing her or his best on the SAT, test centers have been established in nearly every public high school in Maine. The test center supervisors are responsible for planning the use of the facility and selecting rooms with adequate seating, lighting, and ventilation; access to restrooms; and seclusion from noisy areas or distracting activities (e.g., band practice). To discourage copying, all seats in a testing room must face the same direction with at least four feet between each student. No material (e.g., charts, posters) that could be of assistance to a test taker can be displayed in the room.

4.4 Security

Three important facets to the security of a test administration are ensuring that no test taker has had prior access to the content of the test, that the test taker is indeed the person registered for the test, and that the test taker receives no assistance in responding to the test.

The physical security of all testing materials is fundamental to a fair and equitable administration. The SAT test center supervisor is responsible for receiving the test materials, checking them to ensure that they corresponded with what was shipped, and storing the materials in a locked storage area that is not accessible to students or other staff. Test materials are accounted for several times during the day of testing—when the test books and answer sheets are distributed to students, when they are collected from the students, and as they are packed for return to the SAT Program. Supervisors are encouraged to return test materials to the SAT Program immediately after the test, although many may have to be picked up for return shipping to the SAT Program on the Monday following the test (or even later for students whose accommodations required that they be tested in school during the week).

Even though nearly all students test in their own high school, admission to the test center is carefully monitored. Students are instructed to bring their SAT admission ticket and an acceptable photo ID, which are checked against both the admission ticket and the attendance roster previously provided to the supervisor.

Students are not permitted to choose their own seats; rather, they are assigned seating by the supervisory staff to minimize the opportunity for preplanned collaboration among friends. No unauthorized person is permitted to enter the testing room after the administration has begun.

The materials that students may have on their desk during testing are very limited: the test book, answer sheet, No. 2 pencils (pens are not permitted), erasers, and, for the SAT mathematics sections, a calculator. Although all mathematics questions on the SAT can be solved without a calculator, students are encouraged to bring a graphing or scientific calculator. Materials approved as an accommodation for students with disabilities are the only exceptions made to these restrictions.

Test takers are strictly prohibited from using alarm watches or watches containing cameras; protractors; compasses; rulers; dictionaries or other books; pamphlets; papers of any kind; highlighters; colored pens or pencils; recording, copying, or photographic devices; pagers; handheld computers; electronic devices of any type; or cell phones. Handheld computers must be turned off and stored out of sight. Pagers and cell phones were not allowed at the test center. Violation of these prohibitions could lead to dismissal from the testing session and/or cancellation of test scores. Note that computer use is only allowed for students with approved accommodations to use a computer (e.g., to write their essays). This accommodation requires that the student be tested in school using a computer provided by the school.

As a further step to prevent students from helping each other (deliberately or inadvertently), a number of test book variants are used during any one administration. At any given time some students may be working on a mathematics section, some on a critical reading section, and some on a writing section.

4.5 Calculator Policy for the SAT

Calculators are permitted for the entire mathematics section of the SAT. It is recommended that students use a graphing calculator or a scientific calculator. Four-function calculators are not recommended. Every question on the test can be solved without a calculator; however, using a calculator on some questions may be helpful. Students are encouraged to bring a calculator with which they are familiar and should know how and when to use their calculator.

Most calculators, even those with computer algebra systems (CAS) are permitted on the SAT. Unacceptable calculators are those that

- use QWERTY (typewriter-like) keypads;

- require an electronic outlet;
- "talk" or make unusual noises;
- use paper tape; or
- are electronic writing pads, pen input/stylus-driven devices, pocket organizers, cell phones, power books, or handheld laptop computers.

4.6 ITEM TYPES

The mathematics test of the SAT contains two types of questions:

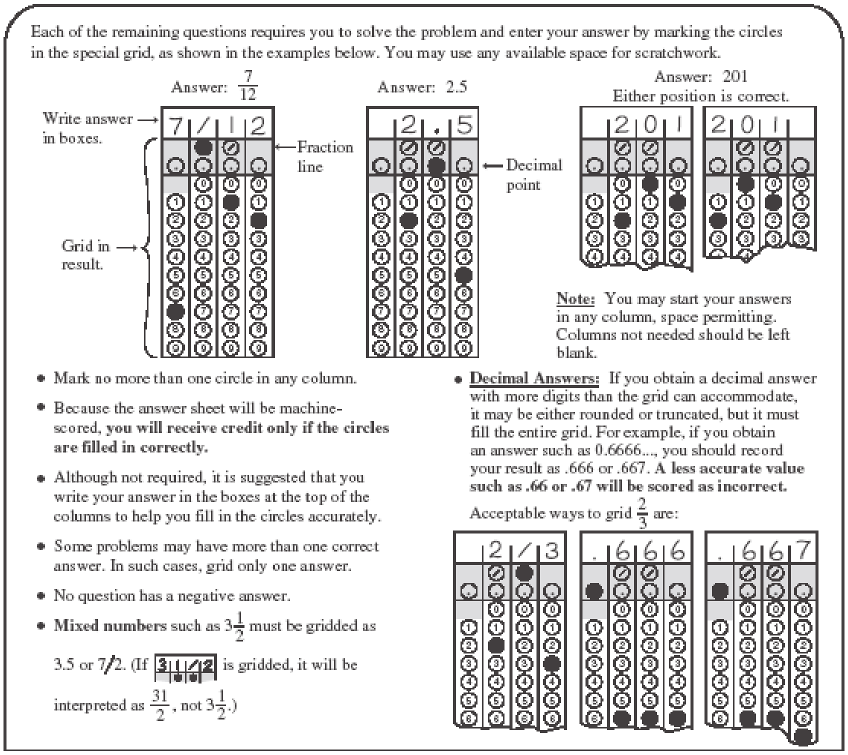
- Standard multiple-choice (44 questions)
- Student-produced-response questions that provide no answer choices (10 questions)

For student-produced-response questions, no answer choices are provided. Students must solve the problem and fill in the answer on a special grid. The directions are fairly simple, and the gridding technique is similar to the way other machine readable information is entered on forms.

A primary advantage of this format is that it allows students to enter the form of the answer that they obtain, whether whole number, decimal, or fraction. For example, a student who obtains an answer of 2/5 can grid 2/5. If a student obtains an answer of 0.4 to the problem, the answer can be gridded in that form as well.

It is virtually impossible to guess an answer to a student-produced-response question, so they are highly reliable. There are no points deducted for incorrect answers to these questions. Table 4-1 shows the actual test directions for student-produced-response items.

Table 4-1. 2013–14 MHSA: SAT Instructions for Student-Produced Responses



Each of the remaining questions requires you to solve the problem and enter your answer by marking the circles in the special grid, as shown in the examples below. You may use any available space for scratchwork.

Answer: $\frac{7}{12}$

Write answer in boxes.

Fraction line

Grid in result.

Answer: 2.5

Decimal point

Answer: 201
Either position is correct.

Note: You may start your answers in any column, space permitting. Columns not needed should be left blank.

- Mark no more than one circle in any column.
- Because the answer sheet will be machine-scored, **you will receive credit only if the circles are filled in correctly.**
- Although not required, it is suggested that you write your answer in the boxes at the top of the columns to help you fill in the circles accurately.
- Some problems may have more than one correct answer. In such cases, grid only one answer.
- No question has a negative answer.
- **Mixed numbers** such as $3\frac{1}{2}$ must be gridded as 3.5 or 7/2. (If 3 1 / 2 is gridded, it will be interpreted as $\frac{31}{2}$, not $3\frac{1}{2}$.)

- **Decimal Answers:** If you obtain a decimal answer with more digits than the grid can accommodate, it may be either rounded or truncated, but it must fill the entire grid. For example, if you obtain an answer such as 0.6666..., you should record your result as .666 or .667. **A less accurate value such as .66 or .67 will be scored as incorrect.**

Acceptable ways to grid $\frac{2}{3}$ are:

2/3 | .666 | .667

4.7 Instructions and Timing

Central to the concept of standardized testing is the notion that all students should receive exactly the same instructions and be given precisely the same amount of time to work on the several parts of a test. To achieve standardization, the SAT Program provides a script for associate supervisors to read and instructions about the amount of time allowed for each of the specific sections of the test. This rule also applies to students receiving extended time as an approved accommodation; they are permitted 50% or 100% additional time for each section of the test, while the room supervisor strictly controls when they start and stop each section.

4.8 Complaints and Irregularities

Because hundreds of people are involved in administering the SAT in Maine, certain situations may not conform to the standardized model. Each irregularity is documented, including any action taken at the test center to remedy the situation. Supervisors are provided with instructions for dealing onsite with many common irregularities. All reports of irregularities are reviewed by Test Administration Services and SAT Program staff to determine whether the occurrence was severe enough to invalidate the test scores of the students involved.

4.9 SUBGROUP PERFORMANCE

In accordance with NCLB legislation that subgroup performance be analyzed and reported, Tables K-3 to K-8 in Appendix K present the number of examinees from Maine in each subgroup along with the mean and standard deviation for each subgroup in mathematics, critical reading, and writing. To protect student confidentiality of test scores, the MDOE does not report mean scores and standard deviations for subgroups containing fewer than five examinees.

4.10 ACCOMMODATIONS FOR STUDENTS ON THE MHSA

Accommodations for students who cannot access state assessments through standard administration are available on the MHSA, as they are for the state assessment in grades 3 through 8. They are designed to allow all students with unique learning needs a fair opportunity to demonstrate what they know and can do at the high school level. The decision to allow the use of accommodations by an individual on any state assessment must be made by the student's IEP or other team of educators.

There are two categories of accommodations for the MHSA: (1) those approved by the College Board through the Eligibility Form process, and (2) those approved only by the State of Maine, designated as Maine Purposes Only (MPO). The accommodations listed for either category are equivalent. In order to assure the opportunity for all Maine students to participate in the SAT component of the MHSA, the College Board agreed to allow some Maine third-year high school students to use accommodations selected from a state approved MPO list, with the understanding that the scores would be used strictly for Maine adequate yearly progress (AYP) purposes and not result in scores reportable to colleges for admission. The same accommodations are included in both categories.

Students with an identified disability are instructed to apply first for College Board approval by submitting a Student Eligibility Form to the College Board. Students may include any MPO accommodations under the category "Other" on the Student Eligibility Form. College Board approval of the accommodations allows students to take the SAT portions of the MHSA and receive college reportable scores. Students, whose accommodations requests have not met College Board criteria or who did not apply for accommodations through the College Board, are still eligible for MPO accommodations if approved by a local district team. For state assessment reporting purposes, there is no difference based on the type of accommodation used. However, only those students using College Board–approved accommodations receive official SAT scores that can be reported to colleges. School personnel are instructed to provide the same accommodations on all components of the MHSA as appropriate.

Historically, about 8.9% of those taking the state-administered MHSA tests have qualified for testing accommodations. Nationally, approximately 2.2% of SAT test takers qualify for College Board approved Services for Students with Disabilities (SSD) accommodations. In the May and June 2014 administrations, 10.5% of those taking the MHSA qualified for testing accommodations: 10.2% in reading, 10.2% in mathematics, and 10.2% in writing.

4.10.1 Process and Standards for College Board–Approved Accommodations

In order to be eligible for College Board approved accommodations, the student must have a documented disability that substantially limits the student’s ability to participate in College Board tests. The College Board Services for Students with Disabilities (SSD) offers two ways for a student to be determined eligible for accommodations on its tests.

1) **School verification:** When a student’s school generated individualized education program (IEP), 504 plan, or other formal written educational plan/program and its supporting documentation align with the College Board’s eligibility criteria and guidelines, and officials at the student’s school verify this to be accurate, the College Board generally does not need further documentation. The College Board processes the form and notifies the student and school of the approved accommodations. To qualify for accommodations under the school verification process, the student must:

    - Have a disability that necessitates testing accommodations

    - Have documentation on file at school that supports the need for the requested accommodation and meets the College Board's Documentation Guidelines

    - Receive and use the requested accommodations, due to the disability, for school-based tests for four school months

    The accommodations request must be submitted by the student’s school, and the school must have an SSD Coordinator form on file with the College Board

2) **Documentation review:** If all of the above requirements are not met, a student may still be eligible for accommodations on College Board tests. The student’s disability documentation is submitted for the College Board’s review and a panel of experts in educating and assessing students with disabilities reviews the documentation and advises the College Board as to whether the documentation supports the request for accommodations. Documentation review is also available for students who want the College Board to make a determination without their school’s involvement. The College Board Guidelines for Documentation require that documentation:

    - state the specific disability, as diagnosed;

    - be current (in most cases, the evaluation and testing should be completed within five years of the request for accommodations). For medical and psychiatric disabilities, an annual evaluation update must be within 12 months of the request for accommodations;

    - provide relevant educational, developmental, and medical history;

- describe the comprehensive testing and techniques used to arrive at the diagnosis, including evaluation date[s] and test results with subtest scores.
- describe the functional limitations (how the disability impacts learning and ability to participate in the test).
- describe the specific accommodations requested, including the amount of extended time required if applicable. State why the disability qualifies the student for such accommodations on standardized tests; and
- establish the professional credentials of the evaluator, including basic information about license or certification and area of specialization.

The guidelines are included in the instructions for the Student Eligibility Form and are also available on the College Board Web site at www.collegeboard.org/ssd.

4.10.2 Process and Standards for MPO Accommodations

Maine has historically allowed testing accommodations to be provided to students, regardless of disability identification, if approved by a local team of educators. As these accommodations are not necessitated by limitations on the ability to participate in College Board tests due to disability, they would not be available on any ordinary, college reportable administration of a College Board test. These accommodations include

- services for students who are limited English proficient (e.g., bilingual dictionaries, word lists); and
- services for "at risk" students who perform poorly under standardized testing conditions but have no identified or suspected disabilities (e.g., extra time).

Maine's state assessment policies and practices allow accommodations for students other than those with disabilities. Such students include those who are ill or incapacitated in some way, those with limited English proficiency, those with a 504 plan, or those for whom classroom accommodations are necessary on a daily basis to measure academic achievement. The "Policies and Procedures for Accommodations and Alternate Assessment" is presented in Appendix B. The MPO accommodations have been designed to be comparable to those available to students approved by the College Board through the Eligibility Form process.

4.10.3 Eligibility Process Additions to Incorporate MPO Accommodations

Maine students with disabilities were encouraged to apply for testing accommodations through the College Board's SSD eligibility process. Maine students who are approved for testing accommodations through the SSD eligibility process are allowed to be tested through existing College Board processes for SSD center-based SAT testing and SSD school-based SAT testing. Tests administered through these processes with approved accommodations are considered valid by the College Board and become part of the student's SAT record maintained by the College Board.

As noted above, Maine students who desire testing accommodations not approved by the SSD eligibility process are allowed to take the test if the additional or alternate accommodations are approved by a local team of Maine educators. Refer to Appendix D for a list of specific MPO accommodations. Under this process, the test is scored by the College Board but is not considered a valid SAT administration and does not become part of the student's SAT record.

MPO accommodations are granted both in cases for which the College Board SSD approved no accommodations and in cases for which the College Board SSD approved fewer accommodations than did an IEP team. In both cases, the student's family and school IEP team are afforded the final decision whether to take the test with the level of accommodations approved by the College Board and have the test applied to the student's SAT record, or to take the test with the MPO accommodations and forfeit the SAT record.

Each Maine high school coordinator is assigned ultimate responsibility by the MDOE for ensuring all students with disabilities are processed through the College Board SSD and Maine-specific eligibility processes (working directly with the designated College Board SSD coordinator and/or Maine eligibility coordinator as necessary).

4.10.4 Accommodation Eligibility Form Submission Time Lines

To assist Maine in organizing its students' requests for accommodation and providing for sufficient time for students to choose between College Board–approved accommodations and MPO accommodations, an earlier submission deadline is established for accommodation eligibility forms to be submitted to the College Board SSD.

Specifically, a February 3, 2014, deadline was established for Maine high school junior eligibility form submissions. The standard deadline for eligibility form submissions for the May 3, 2014 SAT was March 14, 2014.

4.10.5 Training and Technical Assistance

Workshops were conducted by College Board program staff in collaboration with MDOE personnel in order to fully inform individual school representatives about the MHSA and associated deadlines. Rather than conducting separate workshops for issues involving students with disabilities, this information was incorporated into the regularly scheduled training workshops. Workshops were conducted via the Web on February 11, 2014.

4.10.6 MHSA Accommodation Request and Approval Statistics

Table 4-2 presents the numbers of accommodations requested and approved and the types of accommodations approved for Maine public school juniors or third-year students for the 2014 MHSA administration. It includes any approvals for students who chose to take the test under MPO conditions.

Table 4-2. 2013–14 MHSA: Summary of Accommodations for 2013 MHSA Administration

Total number of accommodations requested for College Board approval	*4,489*
Total number of accommodations approved by College Board	4,034
Total number of students using College Board accommodations	1,156
Total number of students using MPO accommodations	107
Total number of students using accommodations	1,263
MPO Accommodations	
MT1–Extended time same day	64
MT2–Extended time over several days	12
MT3–Multiple or frequent breaks	15
MS1–School location other than classroom	2
MS2–Offsite location with school personnel	2
MP1–Individual testing	6
MP2–Small group testing	72
MP3–Human reader	19
MP5–Stand, move, pace during testing	2
MP7–Proctored by special education or ESL Title 1 personnel	36
MP10–Bilingual dictionary	24
MR1–Scribe/recording device for other than essay	7
MR6–Visual Aids	1
MR7–Bilingual dictionary	2
MR8–Verification directions understood	40
MO1-Accommodations based on Test Content	6
Maine Only Accommodations	17
College Board Accommodations	
Large print–20 point	4
Large block answer sheet	8
Braille test	1
Large print-14 point	5
Braille device for written responses	1
Reader	147
Cassette test version	2
Writer to record responses	77
Computer to record written responses	53
Reading–50% extended time	674
Writing–50% extended time	669
Mathematical calculations–50% extended time	655
Listening–50% extended time	23
Speaking – 50% extended time	19
Reading–100% extended time	23
Writing–100% extended time	27
Mathematical calculations–100% extended time	22
Listening – 100% extended time	2
Speaking – 100% extended time	2

continued

Extra breaks	355
College Board Accommodations	
Written directions, or bring sign language interpreter	6
Extended breaks	87
Snacks and/or fluids permitted, medication permitted	15
Preferential seating	26
Write answers in the test book	2
Breaks as needed	34
School-based testing	8
Test blood sugar level	17
Small group setting	800
One-to-one testing	24
Assistive Technology	1
Auditory Amplification/ FM System	1
Other – Other Center Based	3
Other	1

* Students may be granted more than one accommodation and therefore may appear in multiple counts within the table. The listing of accommodations is not comprehensive. Accommodations with counts of 0 were omitted.

4.11 Participation

The intent of the MHSA is for all students in their third year of high school to participate in all components of the test. However, on those occasions where it was necessary to grant a waiver to students from taking the SAT due to special considerations, such as hospitalization or a death in the family, schools were asked to seek the approval of the MDOE MHSA coordinator. The MHSA Operational Procedures document located at http://www.maine.gov/education/mhsa/testmaterial.html describes the criteria for special considerations. Approved students’ nonparticipation was reported in the MHSA results.

CHAPTER 5 THE MHSA SCIENCE COMPONENT: TEST ADMINISTRATION

As the contractor responsible for the administration of the science test, Measured Progress completed tasks such as printing and shipping the test materials, arranging for the return and log-in of test materials, scanning the answer documents, providing an online version of the paper and pencil test, and conducting item analysis for production of student results.

The science test was administered at all Maine high schools during the testing window of March 24 to April 4, 2014. As indicated in the *Principal and Test Coordinator Manual* and the *Online Test Administration Manual*, principals and/or their designated MHSA coordinators were responsible for the proper administration of the science portion of the MHSA. Manuals containing explicit directions and scripts for test administrators to read aloud to test takers were used to ensure the uniformity of administration procedures from school to school.

5.1 RESPONSIBILITY FOR ADMINISTRATION

To ensure the administration of the science test in a fair, equitable, and standardized manner, principals and/or schools' designated MHSA coordinators were instructed to read the *Principal and Test Coordinator Manual* and/or the *Online Test Administration Manual* prior to testing and to be familiar with the instructions given in the *Test Administrator Manual*. The *Principal and Test Coordinator Manual* and the *Online Test Administration Manual* provided checklists to help schools prepare for testing before, during, and after test administration. Along with these checklists, the *Principal and Test Coordinator Manual* outlined the nature of the testing material being sent to each school, how to inventory the material, how to track it during administration, and how to return the material once testing was complete. The *Test Administrator Manuals* also included checklists for administrators to ready themselves, their classrooms, and the students for the administration of the test. The *Test Administrator Manuals* contained sections detailing the procedures to be followed during testing, as well as instructions on preparing the material for its return to Measured Progress. The manuals may be accessed at http://www.maine.gov/doe/mhsa/administration/index.html.

In addition to distributing the *Principal and Test Coordinator Manual, Online Test Administration Manual*, and *Test Administrator Manual*, the MDOE conducted a live and broadcast test administration workshop across the state to train and inform school personnel about the science test. The test coordinator was responsible for the security of the tests while within the schools. Information concerning test security and ethical administration is clearly spelled out in both manuals and stressed during the test administration workshop. Principals were required to complete an online *Principal's Certification of Proper Test Administration* form at the conclusion of testing, certifying that all testing was administered according to

MHSA protocols, verifying the number of students tested either online or on paper, and indicating the number of student response booklets being returned.

5.2 Participation Requirements and Documentation

The intent of the MHSA is for all students in their third year of high school to participate in testing through standard administration, administration with accommodations, and/or alternate assessment. Any student who is absent during the test session is expected to take a makeup test within the testing window.

Eligibility for taking the science test with accommodations was determined during the registration process for the SAT conducted by the College Board. (Please see Chapter 4 for a complete description of this process and a chart showing the numbers of students who tested using accommodations.) School personnel were advised in the *Principal and Test Coordinator Manual* and the *Online Test Administration Manual,* in test administration workshops run by the College Board and the MDOE, and by information posted on the MDOE Web site that students were to take the science test using the same approved accommodations documented during the SAT registration process.

On those occasions when it was necessary to grant a student a waiver from taking the science test due to special considerations, such as hospitalization or a death in the family, schools were asked to seek the approval of the MDOE MHSA coordinator. The names of these students were forwarded to Measured Progress so they would not be included in any reports. A summary of participation rates, both overall and by demographic categories, is provided in Appendix C.

5.3 Test Security

Maintaining test security is critical to the success of the MHSA. The *Principal and Test Coordinator Manual, Online Test Administration Manual,* and *Test Administration Manual* explain in detail all test security measures and test administration procedures. School personnel were informed that any concerns about breaches in test security were to be reported to the school's test coordinator and/or principal immediately. The test coordinator and/or principal were responsible for immediately reporting the concern to the District Superintendent and the State Assessment Director at the MDOE. Test security was also strongly emphasized at test administration workshops. Principals were required to log onto a secure Web site to complete the *Principal's Certification of Proper Test Administration* form; they also had to provide the number of secure tests received from Measured Progress, the number of tests administered to students, the number of students tested online, and the number of secure test materials they were returning to Measured Progress. Principals were instructed to submit the form by entering a unique password, which acted as their digital signature. By signing and submitting the form, the principal certified that the tests were administered according to the test administration procedures outlined in the *Principal and Test Coordinator Manual, Online Test Administration Manual,* and *Test Administration Manual*, that the security of the tests was maintained, that no secure material was duplicated or in any way retained in the school, and that all test materials had been accounted for and returned to Measured Progress.

5.4 Test and Administration Irregularities

There were no test irregularities in the spring 2014 administration.

5.5 Test Administration Window

The test administration window was March 24 through April 4, 2014.

5.6 Service Center

To provide additional support to schools before, during, and after testing, Measured Progress established the MeCAS Service Center and the Measured Progress Technical Product Support Helpdesk. The support of the Service Center is essential to the successful administration of any statewide test program. These service centers provide a centralized location that individuals in the field can call using a toll-free number to ask specific questions or report any problems they may be experiencing. Representatives are responsible for receiving, responding to, and tracking calls and then routing issues to the appropriate person(s) for resolution. All calls are logged into a database, which includes notes regarding the issue and resolution of each call. The Service Center was open to receive calls from 7:30 a.m. to 4:30 p.m. Monday through Friday beginning two weeks before the start of testing and ending two weeks after testing.

CHAPTER 6 SCORING: SAT

Most students, parents, teachers, guidance counselors, and college admissions officers are familiar with the SAT score scale of 200 to 800. This chapter will describe the process of scaling the score.[4] The first portion of the chapter focuses on the process of receiving the completed answer sheets and materials and the associated quality control process; the second portion focuses on the majority of the test—those questions and responses that can be scored by machine; the third portion describes scoring the essay section of the SAT writing test—a process that involves experienced teachers facilitated by electronic technology.

6.1 RECEIVING AND OPENING

Upon completion of the SAT, test center supervisors begin to pack the answer sheets and ancillary materials into shipping cartons with pre-affixed tracking labels. Each test center shipment is routed to the answer sheet processing center in Austin, Texas. The tracking labels are associated with each unique testing center. The tracking labels are scanned, matching them to test centers, which enables the identification of missing or incomplete shipments from the center.

Shipments are then moved into opening, where materials are removed from the shipping cartons. Representatives perform a quality review of the Supervisor Report Form and visually inspect answer sheets for obvious n-count discrepancies. Discrepancies are isolated to the individual test taker and held for resolution. Answer sheets are batched and placed on carts in preparation for scanning.

Ancillary materials are reviewed and forwarded to the applicable departments. Ancillary materials include, but are not limited to, the following:

- Standby registrations
- Cancellation forms
- Supervisor Irregularity Report (SIR)
- Supervisor Report Form (SRF)
- Student Information Correction form
- Seating charts
- Test Question Ambiguity/Error form

[4] Chapter 9 describes how scores are transformed to the MHSA scale of 1100 to 1180.

6.2 Scanning and Editing

Scanning is a single pass operation that captures demographic data, form data, item response data, and essay images from each side of the answer sheet. Answer sheets are held in a climate controlled environment and scanned twice. Discrepant items are reviewed by an editor to determine which scan value should be captured. The following quality controls regulate the scanning process:

- Prior to starting a batch of answer sheet documents on a scanner, the operator must successfully run 10 diagnostic sheets to ensure scanner calibration. The scanner must accurately read 59,220 ovals without an error; the scan program does not proceed unless the diagnostic sheets have been read successfully.
- Prior to the scanning of each batch, the scanner operator performs a multi-sheet test to ensure the scanner halts if two or more sheets pass through at the same time.
- Each answer sheet has anchor points and timing tracks, which ensure it is properly aligned.
- Periodically, answer sheets receive a hand scan accuracy review, ensuring the scan values match the item responses on the answer sheet.
- Quality control check sheets are placed in every stack to ensure the scanner continues to operate correctly.

Additional quality checks at edit include the following:

- Resolve conditions where the information was written but not gridded. Fields include name, social security number, date of birth, gender, and registration number.
- Validate that the test form and form code on the answer sheet match the valid values for the administration date.
- Ensure that only those students with authorized accommodations receive the Student Services with Disabilities test form.

6.3 Matching

Matching is the term applied to the process used to associate a candidate's complete and scanned answer sheet with his or her complete and valid registration. There are three types of matches.

1. Auto matching occurs when a specific set of demographic information from the answer sheet matches exactly to the corresponding information from the candidate's registration with a high confidence interval as specified by quality control. There are 10 such data combinations that can result in a high confidence match. Data elements to be matched include, but are not limited to, registration number, last name, first name, date of birth, and gender.

2. Manual matching occurs when combinations of various data elements exactly match the information from the registration, but one or more major data elements (such as registration number) do not match exactly to the registration data. These cases are reviewed to ensure that the correct match is being made even though some data elements are incongruous.

3. Force matching occurs when a registration is neither high confidence nor low confidence matched and is considered to be in an unmatched status. The College Board investigates all unmatched answer documents. The document stays in an unmatched status until it can be high confidence or low confidence matched to a created registration or the College Board declares the need for a force match. Force matching is necessary because it is possible that incomplete demographic information or major discrepancies between registration and answer sheet data, will prevent an answer sheet from ever being high or low confidence matched. During the course of a College Board investigation, it can be determined that a candidate registration and answer sheet should be matched, but the matching cannot take place within established matching rules. At this point, the College Board performs a force match, or override, to associate the answer sheet with the identified registration. This process is subjected to rigorous quality control oversight.

6.4 Machine-Scored Portions

Except for the essay, all SAT critical reading mathematics (including the student-produced responses), and writing questions are scored by machines. Each student answer sheet is optically scanned and converted to a digital file. These digital files are processed by computer, comparing the student response to each item with the official scoring key to determine the number of questions answered correctly, the number answered incorrectly, and the number omitted.

For all multiple-choice questions (each with five options), each wrong answer results in a deduction of ¼ of a point from the total number of right answers to give the corrected raw score, also known as formula scoring. Formula scores are calculated based on the rights, wrongs, or omits, taking into account the penalty for incorrect responses. For SAT mathematics, the total number right among the student-produced-response questions is added to the corrected raw score for the multiple-choice questions to produce the total raw score. For SAT writing, the corrected raw score for the multiple-choice questions is combined with the essay score to produce the total raw score.

Prior to each administration, a test set of answer sheets consisting of all right and all wrong answers is run through the formula score process. This quality control check is designed to determine if the correct score keys within the system are valid. Upon successful completion of this check, the administration is approved for answer sheet processing.

The raw score for each of the three sections is converted to the 200–800 point score scale through a statistical process called equating. Equating ensures that the varying difficulty levels of different forms of the test do not affect the scaled score that is reported. Equating allows comparisons among test takers who take different editions of the test across different administrations. This process is described in more detail in Chapter 8.

Conversion is a system activity that applies the conversion tables produced during equating to raw formula and essay scores to generate the scaled scores. Conversion quality assurance for each administration includes manually converting a randomly selected, statistically valid sample of answer sheets, through independently generated tables, and comparing the resulting scaled scores to the systematic results produced.

6.5 Scoring the Essay

The SAT essay responses are scored by experienced high school teachers and college faculty members who teach either English or another subject that requires a substantial amount of writing. To be considered for the position of essay reader, a person must

- hold a bachelor’s degree or higher;
- teach or have taught a high school or college level course that requires writing;
- have taught for at least a three-year period;
- reside in the continental United States, Alaska, or Hawaii; and
- be a U.S. citizen, a resident alien, or authorized to work in the U.S.

In addition, readers must complete a rigorous online training course on the principles of holistic scoring that teaches them to evaluate essays according to the agreed-upon standards.

The qualification process, which takes 10 to 15 hours, requires readers to score 30 papers that have previously been scored by leadership and approved by the College Board. To qualify to serve as a reader, a person must score these qualifying papers consistently with leadership, either assigning the same exact score to at least 70% of the papers OR scoring at least 50% exactly, with at least 90% within one point (exact or adjacent).

The pool of readers available for essay scoring is very large, and every effort is made to ensure diversity in terms of gender, ethnicity, education level, and teaching experience. The exact breakdown of rater characteristics for any one administration varies due to demand for and availability of readers. Confidentiality requirements permit readers to omit or choose not to answer some background questions, and therefore the exact percentages in the pool may vary from those reported. The reader pool for a recent large administration was approximately 23% male and 77% female. The ethnic breakdown was approximately 59% White, 1.5% Native American, 2% Asian, 2% Black, 2% Hispanic, 1.5% Pacific Islander, and 32% unspecified. Approximately 76% of the readers held advanced degrees, with 14% of those at the doctoral level. In terms of teaching experience, 27% of readers reported 3 to 5 years at the high school or college level, 28% reported 6 to 10 years, and 45% reported 11 or more years.

Essays are scored in a fair and consistent manner using a holistic approach. A piece of writing is considered as a total work, the whole of which is greater than the sum of its parts. Readers take into account such aspects as complexity of thought, the substantiality of the development, and facility with language. Holistic scoring recognizes that the real merit of a piece of writing cannot be determined by merely adding together the values assigned to such separate factors as word choice, organization, use of evidence, and adherence to the conventions of written English. A reader does not judge a work based on such separate traits but rather on the total impression it creates, with an emphasis on how these separate factors blend together to become the whole piece of writing.

Readers are trained to be mindful of the conditions under which students wrote the essays and to keep a number of guidelines in mind when scoring essays, including the following:

- Use the scoring guide (displayed in Chapter 2) in conjunction with the sample essays selected for training.
- Read quickly to gain an impression of the whole essay.
- Read the entire essay before scoring, and then score immediately.
- Read supportively, looking for and rewarding what is done well rather than what is done badly or omitted.
- Ignore the quality of handwriting.
- Judge an essay by its quality, not by its length.
- Understand that no one aspect of writing (coherence, diction, grammar) is more important than another, and that no aspect of writing is to be ignored.

Each essay is scored independently by two qualified readers on a scale of 1 to 6, with the combined score for both readers ranging from 2 to 12. (An essay not written on the assignment receives a score of 0.) If the two readers' scores differ by more than one point, a third reader scores the essay. During scoring, readers are also asked to be cognizant of special circumstances that may require flagging due to the following alerted condition codes:

- Off topic, unrelated, or suspected cheating
- Cheating—wrong prompt; valid for a different administration
- On topic but similar to essays read before
- Cry for help—response suggests a situation that warrants investigation, such as the possibility of abuse, depression, or contemplation of suicide
- Confidential data—response contains confidential information such as social security numbers, malicious information about another student, etc.

The accuracy and fairness of the readers are evaluated regularly and frequently through a number of processes. Some of these checks are apparent to a reader, while others are embedded in the flow of student papers. For each administration of the SAT essay, readers are trained by scoring a set of pre-scored calibration essays on the topic(s) used for that administration. The calibration papers are used to clarify issues and provide feedback to the readers.

Maintaining scoring accuracy is further supported through the use of prompt specific anchor papers. Sixteen pre-scored essays are selected as anchor papers to represent the full range of performance across all 6 score points that a reader is likely to see on a given prompt. By comparing operational essays to pre-scored anchor papers, readers are able to assign scores on a given prompt with maximum accuracy. To ensure accuracy across prompts as well, anchor papers are selected by consensus agreement of a test development committee during a process known as range finding. Essays are only selected as anchor papers if members of the range-finding committee, a diverse group of secondary and university teachers, unanimously agree that the level of performance of an essay at a score point matches the level expected for essays at the same score point for other prompts. (For example, the range-finding committee works to ensure that an anchor paper at the 3 score point for prompt A demonstrates the same level of performance as a corresponding anchor paper at the 3 score point for prompt B.)

As a further step in maintaining reader accuracy throughout the scoring process, validity papers—clear examples of score points—are interspersed randomly with other student responses. Scoring leaders review readers’ scoring of selected essays and provide feedback via phone and the Web when appropriate. If a reader is unable to accurately score the papers consistently, he or she will not continue as a reader. Web-based scoring enables leaders to monitor readers in real time, informed by extensive real-time and summary reports on interrater reliability, validity, and calibration statistics.

This robust training and monitoring program ensures the highest quality of performance from the readers. As stated previously in this section, a third reading is required when the scores assigned by two readers differ by more than one point. Less than 2% of the 2014 SAT essays from all tests taken in May and June required a third reading (Figure 6-1), confirming that the rigorous training, qualification process, and continuous monitoring of readers is effective. For the Maine-specific population of students who received official score reports, the percentage of essays requiring a third reading was also less than 2% (Figure 6-2).

Essays are scanned and distributed to readers via the Web. By working with readers via the Web, the College Board is able to attract and involve a larger reader pool from across the country than would be possible at a common site.

Figure 6-1. 2013–14 MHSA: Differences in Reader Scores for National Sample in May and June 2014

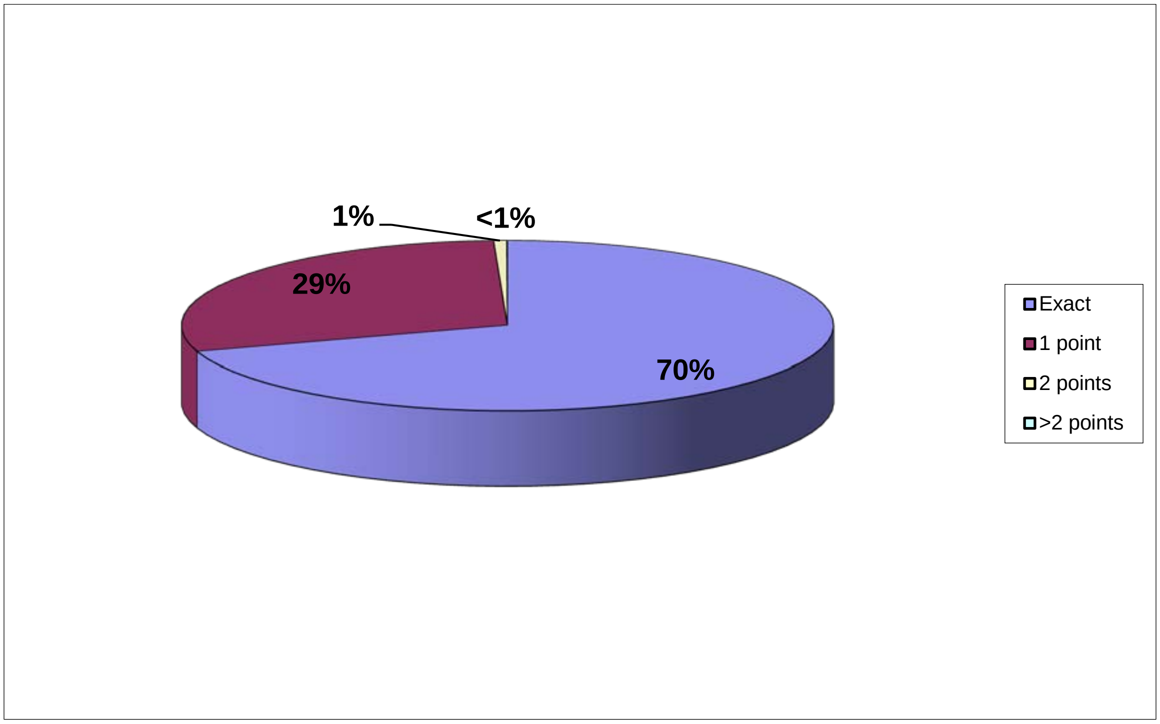
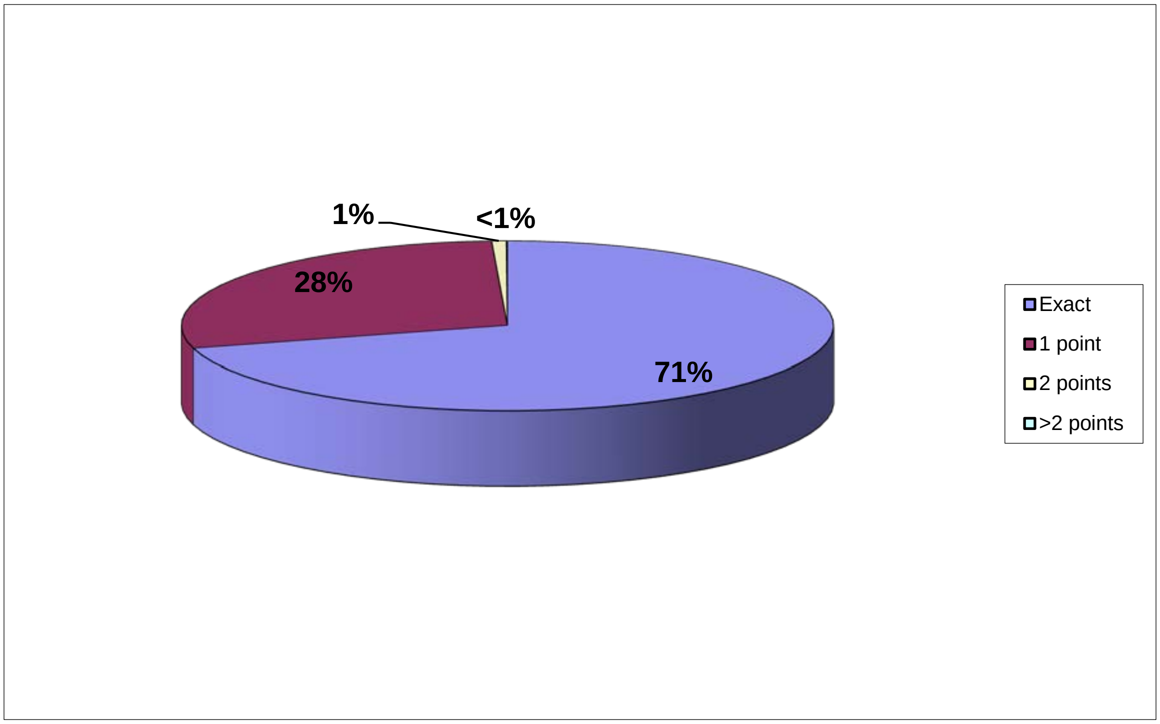


Figure 6-2. 2013–14 MHSA: Differences in Reader Scores for Maine-Specific Sample* in May and June 2014



* Includes data for students receiving official college reportable scores only. Scores for students receiving Maine Purposes Only accommodations cannot be used for college admission or placement purposes.

The scores assigned by the two readers are combined into an essay subscore ranging from 2 to 12. The distribution of scores assigned in the May and June 2014 national administrations for all test takers is shown in Figure 6-3. The Maine-specific distributions for May and June 2014 are displayed in Figure 6-4. It should be noted that Figure 6-4 is based only upon students in Maine who received official College Board score reports for the May and June 2014 administrations.

Figure 6-3. 2013–14 MHSA: National Distribution of SAT Essay Scores for May and June 2014

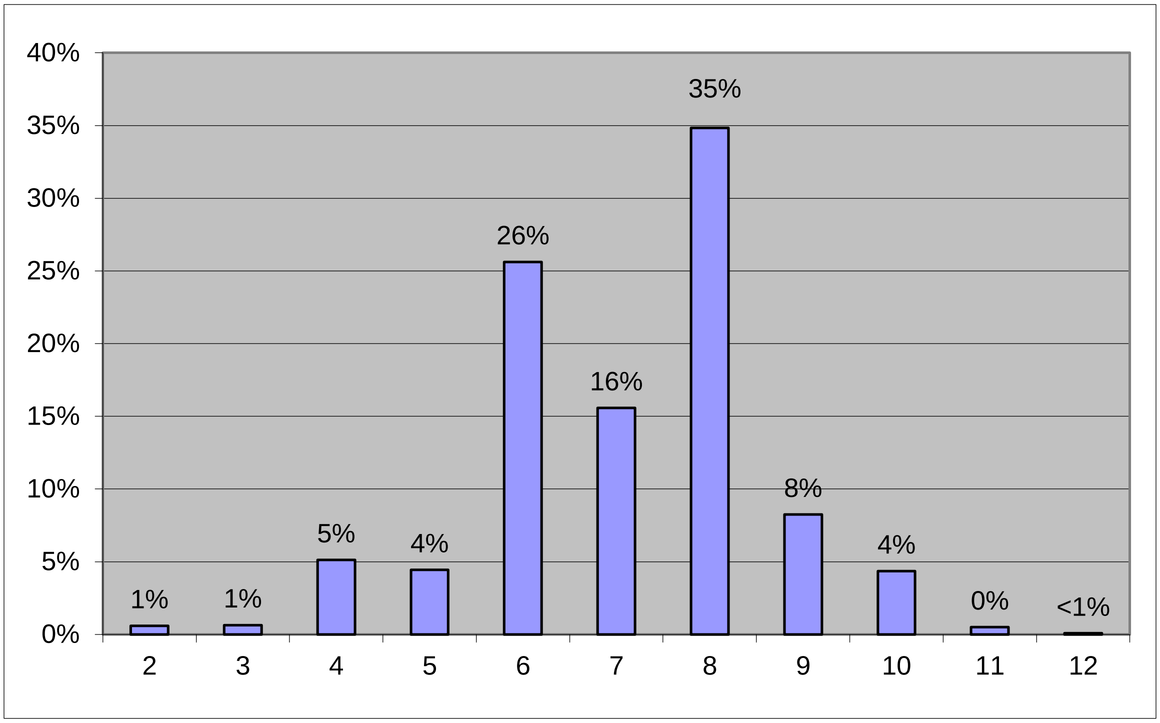
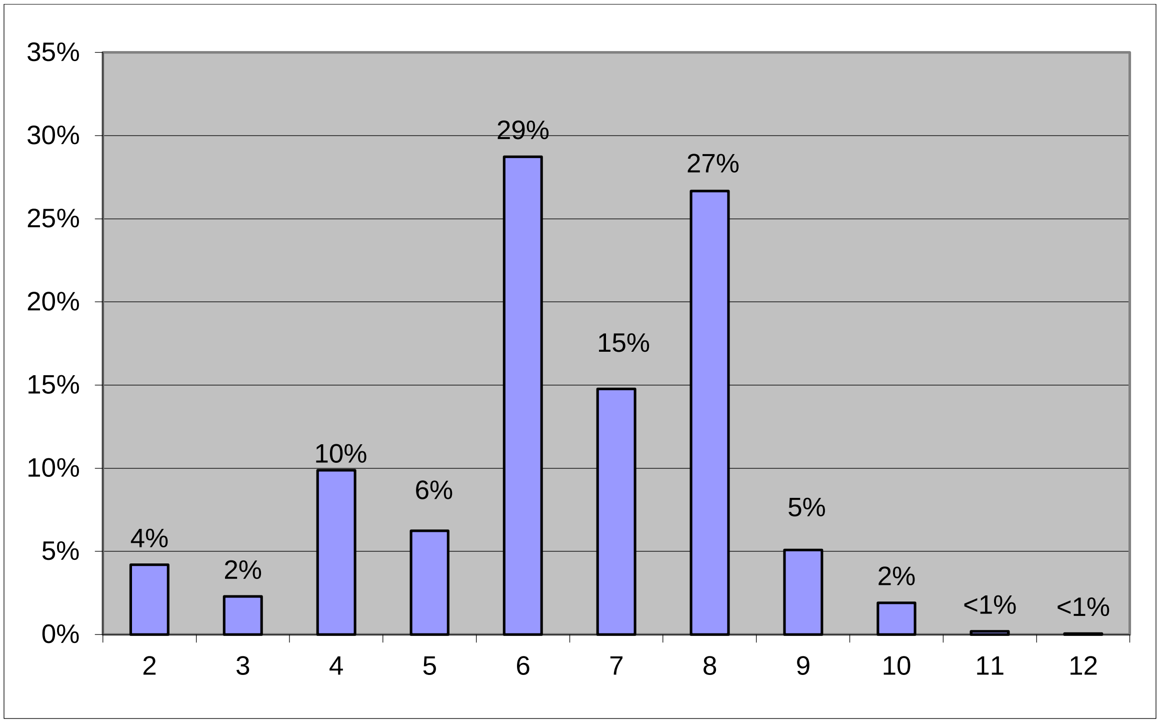


Figure 6-4. 2013–14 MHSA: Maine-Specific Distribution* of SAT Essay Scores for May and June 2014



*Includes data for students receiving official college reportable scores only.

The essay score is combined with the raw score earned on the multiple-choice portion of SAT writing and converted to the 200–800 point scale. The essay score constitutes approximately 30% of the total raw score, and the multiple-choice section makes up the remaining 70%. The distribution of SAT writing scores for the national 2014 College Board college-bound Seniors cohort and the associated percentile ranks are shown in Table 6-1.

Table 6-1. 2013–14 National SAT Writing Percentile Ranks*

Score	Writing Percentile Rank	Score	Writing Percentile Rank	Score	Writing Percentile Rank
800	99+	590	80	380	16
790	99	580	78	370	14
780	99	570	76	360	12
770	99	560	73	350	10
760	99	550	70	340	8
750	98	540	68	330	7
740	98	530	65	320	5
730	97	520	62	310	4
720	97	510	58	300	4
710	96	500	55	290	3
700	96	490	52	280	2
690	95	480	48	270	2
680	94	470	45	260	2
670	93	460	41	250	1
660	92	450	38	240	1
650	90	440	34	230	1
640	89	430	31	220	1
630	88	420	28	210	1
620	86	410	25	200	–
610	84	400	21	**Mean**	487
600	82	390	19	**SD**	115

Based on the 2014 College Bound Seniors Cohort

As a point of reference, for the SAT writing scores from the 2014 college-bound Seniors cohort had a mean of 488 and a standard deviation of 114.

6.6 End-to-End Quality Control

In addition to specific quality checks at each functional step, the College Board has an end-to-end quality assurance program that follows selected cases from receipt through reporting. The program selects answer sheets from all variations of forms to ensure that what was gridded on the answer sheet is accurately represented in the final delivered score report.

6.7 Quality Assessments

Starting with registration and continuing through score reporting, the College Board's quality engineering department performs onsite process reviews to ensure that all documented procedures have been followed. These assessments include reviewing the results of quality control checks, ensuring that the processes are performing as specified.

6.7.1 Summary

The SAT component of the MHSA is scored through a combination of electronic technology and human readers. The resulting raw scores are then converted to the familiar 200–800 point scale using statistical procedures that ensure the comparability of scores across administrations. These steps allow students, parents, teachers, counselors, and admission officers to use the scores as a common yardstick to augment other student information. These SAT component scores are then translated into Maine's 80-point achievement scale used for accountability purposes at all grade levels from three through eight and high school.

CHAPTER 7 SCORING: SCIENCE

7.1 MACHINE-SCORED ITEMS

Multiple-choice item responses were compared to scoring keys using item analysis software. Correct answers were assigned a score of one point, incorrect answers were assigned -1/3 point, and blanks were zero points. Student responses with multiple marks and blank responses were also assigned zero points.

The hardware elements of the scanners monitor themselves continuously for correct read, and the software that drives these scanners also monitors correct data reads. Standard checks include recognition of a sheet that does not belong or is upside down or backward and identification of critical data that are missing (e.g., a student ID number), test forms that are out of range or missing, and page or document sequence errors. When a problem is detected, the scanner stops and displays an error message directing the operator to investigate and correct the situation.

7.2 PERSON-SCORED ITEMS

The images of student responses to constructed-response items were hand-scored through Measured Progress's electronic scoring system, iScore. Use of iScore minimizes the need for readers to physically handle answer booklets and related scoring materials. Student confidentiality was easily maintained, since all MeCAS scoring was "blind" (i.e., district, school, and student names were not visible to readers). The iScore system maintained the linkage between the student response images and their associated test booklet numbers. Through iScore, qualified readers at computer terminals accessed electronically scanned images of student responses. Readers evaluated each response and recorded each score via keypad or mouse entry through the iScore system. When a reader finished one response, the next response appeared immediately on the computer screen.

Imaged responses from all answer booklets were sorted into item-specific groups for scoring purposes. Readers reviewed responses from only one item at a time; however, imaged responses from a student's entire booklet were always available for viewing when necessary, and the physical booklet was also available to the chief reader on-site. (Chief reader and other scoring roles are described in the section that follows.)

The use of iScore also helped ensure that access to student response images was limited to only those who were scoring or working for Measured Progress in a scoring management capacity.

7.2.1 Scoring Location and Staff

Scoring Location

The iScore database, its operation, and its administrative controls are all based in Dover, New Hampshire, which is where all 2013–14 MHSA science test items were scored. The iScore system monitored accuracy, reliability, and consistency across the scoring site. Constant daily communication and coordination were accomplished through e-mail, telephone, faxes, and secure Web sites to ensure that critical information and scoring modifications were shared and implemented across the scoring site.

Staff Positions

The following staff members were involved with scoring the 2013–14 MeCAS responses:

- The MeCAS scoring project manager, an employee of Measured Progress, was located in Dover, New Hampshire, and oversaw communication and coordination of scoring across the scoring site.
- The iScore operational manager and iScore administrators, employees of Measured Progress, were located in Dover, New Hampshire, and coordinated technical communication across the scoring site.
- A chief reader in the science content area ensured consistency of scoring across the scoring site for all grades tested in that content area. Chief readers also provided read-behind activities (defined in a later section) for quality assurance coordinators (QACs). Chief readers are employees of Measured Progress.
- QACs, selected from a pool of experienced senior readers (SRs) for their ability to score accurately and to instruct and train readers, participated in benchmarking activities for each specific grade of the science content area. QACs provided read-behind activities (defined in a later section) for SRs at the scoring site. The ratio of QACs and SRs to readers was approximately 1:11.
- SRs, selected from a pool of skilled and experienced readers, provided read-behind activities (defined in a later section) for the readers at their scoring tables (2–12 readers at each table). The ratio of QACs and SRs to readers was approximately 1:11.
- Readers at the Dover, New Hampshire, scoring site scored operational and field-test MeCAS 2013–14 student responses. Recruitment of readers is described in Section 7.2.3.

7.2.2 Benchmarking Meetings

In preparation for implementing MeCAS scoring guidelines, Measured Progress scoring staff prepared and facilitated benchmarking meetings held with the MeCAS state science specialist representing the department of education. The purpose of these meetings was to establish guidelines for scoring MeCAS items during the current field-test scoring session and for future operational scoring sessions.

Chief readers selected several dozen student responses for each item that were identified as illustrative midrange examples of the respective score points. Chief readers presented these responses to the MeCAS science content specialist during benchmarking meetings and worked collaboratively with him or her to finalize an authoritative set of score-point exemplars for each field-test item. As a matter of practice, these sets are included in the scoring training materials each time an item is administered.

This repeated use of MeCAS-approved sets of midrange score-point exemplars helps ensure that readers follow established guidelines each time a particular MeCAS item is scored.

7.2.3 Reader Recruitment and Qualifications

For scoring the 2013–14 MeCAS, Measured Progress actively sought a diverse scoring pool. The broad range of reader backgrounds typically includes scientists, editors, business professionals, authors, teachers, graduate school students, and retired educators. Demographic information about readers (e.g., gender, race, educational background) was electronically captured for reporting.

Readers were required to have successfully attained a four-year college degree or higher. In all cases, potential readers were required to submit documentation (e.g., résumé and/or transcripts) of their qualifications.

Table 7-1 summarizes the qualifications of the 2013–14 MeCAS scoring leadership and readers.

Table 7-1. 2013–14 MHSA Science: Qualifications of Scoring Leadership and Readers—Spring Administration

Scoring responsibility	*Educational credentials*				*Total*
	Doctorate	*Master's*	*Bachelor's*	*Other*	
Scoring leadership	20.0%	20.0%	60.0%	0.0%	100%
Readers	9.1%	33.3%	57.6%	0.0%	100%

Scoring leadership = chief readers, quality assurance coordinators, and senior readers

Readers either were temporary Measured Progress employees or were secured through temporary employment agencies. All readers were required to sign a nondisclosure/confidentiality agreement.

7.2.4 Methodology for Scoring Polytomous Items

Possible Score Points

The ranges of possible score points for the different polytomous items are shown in Table 7-2.

Table 7-2. 2013–14 MHSA Science: Possible Score Points for Polytomous Item Types

Polytomous item type	Possible score point range
Constructed-response	0–4
Nonscorable items	0

Nonscorable Items

Readers could designate a response as nonscorable for any of the following reasons:

- Response was blank (no attempt to respond to the question).
- Response was unreadable (illegible, too faint to see, or only partially legible/visible)—*see note below.*
- Response was written in the wrong location (seemed to be a legitimate answer to a different question)—*see note below.*

Note: “Unreadable” and “wrong location” responses were eventually resolved by researching the actual answer document (electronic copy or hard copy, as needed) to identify the correct location (in the answer document) or to more closely examine the response and then assign a score.

Scoring Procedures

Scoring procedures for polytomous items included both single scoring and double-blind scoring. Single-scored items were scored by one reader. Double-blind-scored items were scored independently by two readers whose scores were tracked for “interrater agreement” (for further discussion of double-blind scoring and interrater agreement, see Section 7.2.7 and Appendix M).

7.2.5 Reader Training

Reader training began with an introduction of the on-site scoring staff and an overview of the MeCAS program’s purpose and goals (including discussion about the security, confidentiality, and proprietary nature of testing materials, scoring materials, and procedures).

Next, readers thoroughly reviewed and discussed the scoring guide for each item to be scored. Each item-specific scoring guide included the item itself and score-point descriptions.

Following review of an item’s scoring guide, readers reviewed the particular response set organized for that training: Anchor Sets, Training Sets, and Qualifying Sets. (These are defined below.)

During training, readers could highlight or mark hard copies of the Anchor and Training Sets (as well as the first Qualifying Sets after the qualification round), even if all or part of the set was also presented online via computer.

Anchor Set

Readers first reviewed an Anchor Set of exemplary responses for an item. This is a set approved by the science content specialist representing the MDOE. Responses in Anchor Sets are typical, rather than unusual or uncommon; solid, rather than controversial or borderline; and true, meaning that they had scores that could not be changed by anyone other than the MeCAS client and Measured Progress scoring services staff. Each contains one client-approved sample response per score point considered to be a midrange exemplar; each of these responses has, where necessary, the MeCAS science content specialist's rationale for choosing that response as a score-point anchor. The set includes a second sample response if there is more than one plausible way to illustrate the merits and intent of a score point.

Responses were read aloud to the room of readers in descending score order. Announcing the true score of each anchor response, trainers facilitated group discussion of responses in relation to score-point descriptions to help readers internalize the typical characteristics of score points.

This Anchor Set continued to serve as a reference for readers as they went on to calibration, scoring, and recalibration activities for that item.

Training Set

Next, readers practiced applying the scoring guide and anchors to responses in the Training Set. The Training Set typically included 10 to 15 student responses designed to help establish both the full score-point range and the range of possible responses within each score point. The Training Set often included unusual responses that were less clear or solid (e.g., shorter than normal, employing atypical approaches, simultaneously containing very low and very high attributes, and written in ways difficult to decipher). Responses in the Training Set were presented in randomized score-point order.

After readers independently read and scored a Training Set response, trainers would poll readers or use online training system reports to record the initial range of scores. Trainers then led group discussions of one or two responses, directing reader attention to difficult scoring issues (e.g., the borderline between two score points). Throughout the process, trainers modeled how to discuss scores by referring to the Anchor Set and to scoring guides.

Qualifying Set

After the Training Set had been completed, readers were required to score responses accurately and reliably in Qualifying Sets assembled for constructed-response items. The 10 responses in each Qualifying Set were selected from an array of responses that clearly illustrated the range of score points for that item as reviewed and approved by the state specialist. Hard copies of the responses were also made available to readers after the qualification round so that they could make notes and refer back during the post-qualifying discussion.

To be eligible to live score one of the items in the Qualifying Set, readers were required to demonstrate scoring accuracy rates of at least 80% exact agreement (i.e., to exactly match the predetermined score on at least 8 of the 10 responses) and at least 90% exact or adjacent agreement (i.e., to exactly match or be within one score point of the predetermined score on 9 or 10 of the 10 responses). In other words, readers were allowed one discrepant score (i.e., 1 score of 10 that was more than one score point from the predetermined score) provided they had at least eight exact scores.

Retraining

Readers who did not pass the first Qualifying Set were retrained as a group by reviewing their performance with scoring leadership and then scoring a second Qualifying Set of responses. If they achieved the required accuracy rate on the second Qualifying Set, they were allowed to score operational responses.

Readers who did not achieve the required scoring accuracy rates on the second Qualifying Set were not allowed to score responses for that item. Instead, they either began training on a different item or were dismissed from scoring for that day.

7.2.6 Leadership Training

QACs and select SRs were trained in a separate training session immediately prior to reader training. In addition to discussing the items and their responses, QAC and SR training included greater detail on the client's rationale behind the score points than that covered with regular readers in order to better equip QACs and SRs to handle questions from the latter.

7.2.7 Monitoring of Scoring Quality Control

Readers were monitored for continued accuracy and consistency throughout the scoring process, using the following methods and tools (which are defined in this section):

- Embedded Committee-Reviewed Responses (CRRs)
- Read-Behind Procedures
- Double-Blind Scoring
- Recalibration Sets
- Scoring Reports

Note that if a reader's accuracy rate fell below the expected rate for a particular item and monitoring method, the reader was retrained on the item. Upon approval by the QAC or chief reader, as appropriate (see below), the reader was allowed to resume scoring. Readers who met or exceeded the expected accuracy rates continued scoring.

Furthermore, the accuracy rate required of a reader to *qualify* to score live was higher than that required to *continue* to score responses live. The reason for the difference is that an "exact score" in double-blind scoring requires that two readers choose the same score for potentially borderline responses (in other words, is dependent upon two peers agreeing on responses that often do not sit neatly in the middle of the score-point spectrum), whereas an "exact score" in qualification requires only that a single reader match a score pre-established as sitting in the middle of the respective score point by scoring leadership. The use of multiple monitoring techniques is critical to monitoring reader accuracy during the process of live scoring.

Embedded Committee-Reviewed Responses (CRRs)

CRRs are previously scored responses that are loaded (“embedded”) by scoring leadership into iScore and distributed “blindly” to readers during scoring. Embedded CRRs may be chosen either before or during scoring and are inserted into the scoring queue so that they appear the same as all other live student responses.

Between 5 and 30 embedded CRRs were distributed at random points throughout the first full day of scoring to ensure that readers were sufficiently calibrated at the beginning of the scoring period. Individual readers often received up to 20 embedded CRRs within the first 100 responses scored and up to 10 additional responses within the next 100 responses scored on that first day.

Any reader who fell below the required scoring accuracy rate was retrained before being allowed by the QAC to continue scoring. Once allowed to resume scoring, scoring leadership carefully monitored these readers by increasing the number of read-behinds (defined in the next section).

Embedded CRRs were employed for all constructed-response items.

Read-Behind Procedures

Read-behind scoring refers to scoring leadership (usually an SR) scoring a response after a reader has already scored the response. The practice was applied to all constructed-response item types.

Responses placed into the read-behind queue were randomly selected by scoring leadership; readers were not aware which of their responses would be reviewed by their SR. The iScore system allowed one, two, or three responses per reader to be placed into the read-behind queue at a time.

The SR entered his or her score into iScore before being allowed to see the reader’s score. The SR then compared the two scores, and the score of record (i.e., the reported score) was determined as follows:

- If there was exact agreement between the scores, no action was necessary; the regular reader’s score remained.
- If the scores were adjacent (i.e., differed by one point), the SR’s score became the score of record. (A significant number of adjacent scores for a reader triggered an individual scoring consultation with the SR, after which the QAC determined whether or when the reader could resume scoring.)
- If the scores were discrepant (i.e., differed by more than one point), the SR’s score became the score of record. (This triggered an individual consultation with the SR, after which the QAC determined whether or when the reader could resume scoring on that item.)

Table 7-3 illustrates how scores were resolved by read-behind.

Table 7-3. 2013–14 MHSA Science: Examples of Read-Behind Scoring Resolutions

Reader score	*QAC/SR score*	*Score of record*
4	4	4
4	3	3*
4	2	2*

* QAC/SR's score

SRs were tasked with conducting, on average, five read-behinds per reader throughout each half-day of scoring; however, SRs conducted a proportionally greater number of read-behinds for readers who seemed to be struggling to maintain, or who fell below, accuracy standards.

In addition to regular read-behinds, scoring leadership could choose to do read-behinds on any reader at any point during the scoring process to gain an immediate, real-time "snapshot" of a reader's accuracy.

Double-Blind Scoring

Double-blind scoring refers to two readers independently scoring a response without knowing whether the response will be double-blind scored. The practice was applied to all constructed-response item types. Table 7-4 shows by which method(s) both common and equating constructed-response item responses for each operational test were scored.

Table 7-4. 2013–14 MHSA Science: Frequency of Double-Blind Scoring

Grade	*Content area*	*Responses double-blind scored*
HS	Science	10%
HS	Unreadable responses	100%
HS	Blank responses	100%

If there was a discrepancy (a difference greater than one score point) between double-blind scores, the response was placed into an arbitration queue. Arbitration responses were reviewed by scoring leadership (SR or QAC) without knowledge of the two readers' scores. Scoring leadership assigned the final score. Appendix M provides the MeCAS 2013–14 percentages of agreement between readers for each common item for each grade.

Scoring leadership consulted individually with any reader whose scoring rate fell below the required accuracy rate, and the QAC determined whether or when the reader could resume scoring on that item. Once the reader was allowed to resume scoring, scoring leadership carefully monitored the reader's accuracy by increasing the number of read-behinds.

Recalibration Sets

In order for scoring leadership to determine whether readers were still calibrated to the scoring standard, readers were required to take an online Recalibration Set at the start and midpoint of the shift upon their resumption of scoring (daytime shifts are typically 7.5 hours and evening shifts 5.5 hours in duration).

Each Recalibration Set consisted of five responses representing the entire range of possible scores, including some with a score point of 0.

- Readers who were discrepant on two of five responses of the first Recalibration Set, or were exact on two or fewer, were not permitted to score on that item that day and were either assigned to a different item or dismissed for the day.
- Readers who were discrepant on only one of five responses of the first Recalibration Set, and/or exact on three, were retrained by their SR by discussing the Recalibration Set responses in terms of the score-point descriptions and the original Anchor Set. After this retraining, such readers began scoring operational responses under the proviso that the reader's scores for that day and that item would be kept only if the reader was exact on all five of five responses of the second Recalibration Set administered at the shift midpoint. The QAC determined whether or when these readers had received enough retraining to resume scoring operational responses. Scoring leadership also carefully monitored the accuracy of such readers by significantly increasing the number of their read-behinds.
- Readers who were not discrepant on any response of the first Recalibration Set, and exact on at least four, were allowed to begin scoring operational responses immediately, under the proviso that this recalibration performance would be combined with that of the second Recalibration Set administered at the shift midpoint.

The results of both Recalibration Sets were combined with the expectation that readers would have achieved an overall 80% exact and 90% adjacent standard for that item for that day.

The scoring project manager voided all scores posted on that item for that day by readers who did not meet the accuracy requirement. Responses associated with voided scores were reset and redistributed to readers with demonstrated accuracy for that item.

Recalibration Sets were employed for all constructed-response items and were first administered at the start of the second day of scoring on an item, since the first day of scoring an item is monitored using the item's initial qualification set and set of embedded CRRs. In the event an item was scored during a third day, newly assembled Recalibration Sets were administered similarly to how the sets were administered on the second day.

Scoring Reports

Measured Progress's electronic scoring software, iScore, generated multiple reports that were used by scoring leadership to measure and monitor readers for scoring accuracy, consistency, and productivity. These reports are further discussed in the following section.

7.2.8 Reports Generated During Scoring

Because of the complexity of scoring a large-scale assessment project such as that for MeCAS, computer-generated reports were necessary to ensure that

- overall group-level accuracy, consistency, and reliability of scoring were maintained at acceptable levels;
- immediate, real-time individual reader data were available to allow early intervention when necessary; and
- scoring schedules were maintained.

The following reports were produced by iScore for internal use throughout each scoring day by scoring leadership (including SRs, QACs, chief readers, and the scoring project manager, where applicable):

- **The Read-Behind Summary** showed the total number of read-behind responses for each reader and noted the number and percentages of exact, adjacent, and discrepant scores with the SR/QAC. Scoring leadership could choose to generate this report by choosing options (such as “Today,” “Past Week,” and “Cumulative”) from a pull-down menu. The report could also be filtered to select data for a particular item or across all items. This report was used in conjunction with other reports to determine whether a reader’s scores would be voided (i.e., sent back out to the floor to be rescored by other readers). The benefit of this report is that it can reveal the degree to which an individual reader agrees with his or her QAC or SR on how best to score live responses.
- **The Double-Blind Summary** showed the total number of double-scored responses of each reader and noted the number and percentages of exact, adjacent, and discrepant scores with second readers. This report was used in conjunction with other reports to determine whether a reader’s scores should be voided (i.e., sent back out to the floor to be rescored by other readers). The benefit of this report is that it can reveal the degree to which readers are in agreement with each other about how best to score live responses.
- **The Accuracy Summary** combined read-behind and double-blind data, showing the total number of responses scored for the readers, their accuracy rates, and their score-point distributions.
- **The Embedded CRR Summary** showed, for each reader (by item or across all items), the total number of responses scored, the number of embedded CRRs scored, and the numbers and percentages of exact, adjacent, and discrepant scores with the chief reader. This report was used in conjunction with other reports to determine whether a reader’s scores should be voided (i.e., sent back out to the floor to be rescored by other readers). The benefit of this report is that it can reveal the degree to which an individual reader agrees with his or her chief reader on how to best score live responses. Also, since embedded CRRs are administered during the first hours of scoring, this report can provide an early illustration of agreement between readers and chief readers.
- **The Qualification Statistics Summary** listed each reader by name and ID number and identified which Qualifying Set(s) the reader did and did not take and the reader’s pass rate for the sets taken. In addition to the pass rates of individuals, the report also showed numbers

of readers passing or failing a particular Qualifying Set. The QACs could use this report to determine how readers within their scoring group performed on specific Qualifying Sets.

- **The Summary Statistics Report** showed the total number of student responses for an item, and identified, for the time at which the report was generated, the following:
    - the number of single and double-blind scorings that had been performed
    - the number of single and double-blind scorings yet to be performed

CHAPTER 8 PSYCHOMETRIC TOPICS: SAT

The use of the SAT supports Maine's vision of graduating all high school students as college, career, and citizenship ready by assessing how students apply what they have learned in high school to analyze and solve problems they will likely encounter in college. The critical reading section provides a strong focus on the construct of reading, with approximately 72% reading comprehension items. Examinees are allotted 70 minutes to answer the 67 multiple-choice items in the critical reading section. The SAT mathematics section contains 54 items in total—44 multiple-choice and 10 student-produced responses—with an allotted time of 70 minutes to answer the items. The mathematics section covers mathematical concepts through third-year college preparatory mathematics. The writing section contains 49 multiple-choice questions with an allotted time of 60 minutes and a 25-minute section in which the student produces a response to an essay prompt. The writing section is intended to measure how well students use standard written English.

8.1 THE EQUATING AND BRAIDING PLAN FOR SAT MATHEMATICS, CRITICAL READING, AND WRITING

This section outlines the equating and braiding plan for the SAT forms. Equating refers to the statistical process used to ensure that the reported scores on each version of the SAT have the same meaning as every other version. SAT equating employs two types of data collection: the nonequivalent groups anchor test (NEAT) design and the equivalent groups (EG) design. At each SAT administration of one new form, the new form is linked to multiple old SAT forms through a NEAT design. One of the old forms was administered to a similar sample from a similar population—that is, to a sample of students who were administered the SAT during the same month in a previous year. Each of the other old forms was administered at one of the core administrations of the SAT that contribute large numbers of scores to the SAT cohort. The final conversion line is the weighted average line of the four individual lines, with more weight (usually 50%) given to the link to the old form that was administered to a sample from the similar population, defined as the group of students testing in the same administration one year previously. This data collection design has been shown to produce stable equating results because it directly acknowledges the important role that the old form linking plays in placing a new form on scale (Dorans, Liu, and Hammond, 2004).

An EG design is usually employed in an SAT administration with two or more new forms, where the first new form is equated using the NEAT design and the second new form is equated to the first one through an EG design. The spiraling procedure used in the SAT administration and the large numbers of test takers who take each form usually ensure equivalent groups in the same administration.

8.2 SAT STATISTICAL CHARACTERISTICS

The statistical characteristics of the SAT, based on the two forms administered in May and June 2014, are examined in this section. The test-level statistics include reliability, standard errors of measurement (SEM), and test speededness. The item-level statistics include item difficulty, item discriminating power, and differential item functioning (DIF). Analyses for the SAT conducted on the national SAT population and not specific to Maine are presented in Appendix D. Tables D-1 through D-3 provide summaries of the scores for examinees participating in SAT testing in May and June 2014 by section for each form. Tables D-4 through D-6 present the rounded scaled score conversion tables by section for each SAT form.

8.3 RELIABILITY AND STANDARD ERRORS OF MEASUREMENT

8.3.1 Reliability

Reliability is an indicator of the consistency or stability of test scores. Test scores that are used for making important decisions should be very reliable. The estimates of reliability detailed in this report are internal consistency measures, which are derived from analysis of the consistency of the performance of individuals on items within a test (internal consistency reliability). Therefore, they apply only to the test form being analyzed. They do not take into account form-to-form variation due to equating limitations or lack of parallelism, nor are they responsive to day-to-day variation due, for example, to the test taker's state of health or testing environment.

The reliability and SEM on the national equating sample for the mathematics, critical reading, and writing sections are within normally acceptable ranges (see Table D-7 of Appendix D). Due to makeup testing administrations and special forms for students with disabilities, students in Maine took one of four test forms. Using recommendations in the literature as to the size of the sample needed to obtain stable estimates, reliability estimates were calculated only for test forms and subgroups with at least 200 examinees (Kline, 1986; Charter, 1999). The reliability estimates for Maine students only are reported in Table 8-1. These values range from 0.76 to 0.93 for critical reading, 0.80 to 0.93 for mathematics, and 0.73 to 0.89 for writing. This supports the use of SAT scores for students in Maine and is evidence that the reliability of scores for Maine students is comparable to that of the national sample. Reliability estimates were also computed for subgroups that met the minimum sample size requirements: males, females, students with disabilities, students who are economically disadvantaged, and students with limited English proficiency (beyond the first year). Maine subgroup reliabilities are reported in Table 8-2. Subgroup reliabilities range from 0.80 to 0.93 in critical reading, from 0.87 to 0.94 in mathematics, and from 0.75 to 0.89 in writing, with students with disabilities and students categorized as LEP generally showing the lowest reliability coefficients. Average SAT scores and standard deviations on the raw score scale for Maine students are reported in Table 8-3.

8.3.2 Standard Errors of Measurement

The standard error of measurement (SEM) is an estimate of the amount of variation that can be expected in obtained scores for the same individual if the person were to retake the test with no change in knowledge between administrations or for individuals with the same true score. The interpretation of the SEM is usually made in terms of a statement of probability that the score obtained by an individual is within a certain distance of his or her true score (that is, the score he or she would obtain on a perfectly reliable test). The probability is 0.68 that an individual's score will be within one SEM of his or her true score and 0.95 that it will be within two SEMs (assuming a normal distribution). The SEMs for Maine students only are reported in Tables 8-1 and 8-2 for the total Maine group and Maine subgroups, respectively. All raw score SEMs for the total Maine group and for the Maine subgroups ranged from 2.2 to 4.2 for critical reading, 1.8 to 3.4 for mathematics, and 1.9 to 3.7 for writing. Form 2 reliabilities and SEMs were not provided for the Maine-specific sample due to small sample size.

Conditional SEMs (i.e., SEMs at each scaled-score point) are provided in the raw score to scaled score lookup tables, which are presented in Appendix J. These are the actual tables that were used to determine student scaled scores, error bands, and achievement levels.

Table 8-1. 2013–14 MHSA: SAT Reliability Coefficients and SEMs[1] for Sections of the MHSA[2]

		Form	Form 1	
		Administration	05/13	
		Sample N	11,850	
Test Section			*Rel.*	*SEM*
Critical Reading 1	Dressel-KR20	Raw	.82	2.4
Critical Reading 2	Dressel-KR20	Raw	.84	2.5
Critical Reading 3	Dressel-KR20	Raw	.76	2.2
Total Critical Reading	Dressel-KR20[3]	Raw	.93	4.2
	Var. Components	Raw	.93	4.2
Math 1	Dressel-KR20	Raw	.82	2.1
Math 2	Dressel-KR20	Raw	.85	1.8
Math 3	Dressel-KR20	Raw	.80	1.9
Total Mathematics	Dressel-KR20[3]	Raw	.93	3.4
	Var. Components	Raw	.93	3.4
Writing 1	Dressel-KR20	Raw	.85	3.0
Writing 2	Dressel-KR20	Raw	.73	1.9
Total Writing MC	Dressel-KR20[3]	Raw	.89	3.6
	Var. Components	Raw	.89	3.6

[1] See Appendix D for formulas used to compute reliability coefficients and SEMs.
[2] Estimates are computed based on Maine students only for the form that were taken by the majority of Maine students and had sufficient sample size.
[3] Prior to the 2010-2011 year the total section score reliabilities were computed using alpha rather than Dressel-KR20 so some change in the values may be noticed when comparing to earlier manuals.
MC = multiple-choice

Table 8-2. 2013–14 MHSA: SAT Reliability Coefficients and SEMs for Sections of the MHSA*

Test Section	*Subgroup*	*From 1—May 2014*				
		N	*KR-20*		*Variance Components*	
			Reliability	*SEM*	*Reliability*	*SEM*
Total critical reading	Male	5,976	0.93	4.2	0.93	4.2
	Female	5,813	0.92	4.2	0.92	4.1
	Students with disabilities	1,158	0.89	4.2	0.89	4.2
	Economically disadvantaged	4.040	0.91	4.2	0.91	4.2
	Limited English Proficient – Currently Receiving LEP	198	0.81	4.0	0.81	4.0
	Limited English Proficient – Formerly Received LEP	94	0.80	4.1	0.80	4.1
Total mathematics	Male	5,976	0.94	3.4	0.94	3.4
	Female	5,813	0.93	3.4	0.93	3.4
	Students with disabilities	1,158	0.87	3.4	0.87	3.4
	Economically disadvantaged	4.040	0.91	3.4	0.91	3.4
	Limited English Proficient – Currently Receiving LEP	198	0.90	3.3	0.90	3.3
	Limited English Proficient – Formerly Received LEP	94	0.90	3.2	0.90	3.2
Total writing MC	Male	5,976	0.89	3.6	0.89	3.6
	Female	5,813	0.89	3.6	0.89	3.6
	Students with disabilities	1,158	0.81	3.7	0.81	3.7
	Economically disadvantaged	4.040	0.86	3.7	0.86	3.7
	Limited English Proficient – Currently Receiving LEP	198	0.79	3.5	0.79	3.5
	Limited English Proficient – Formerly Received LEP	94	0.75	3.5	0.75	3.5

* Estimates are calculated based on Maine students only for subgroups where sufficient sample sizes were present.

MC = multiple-choice

Table 8-3. 2013–14 MHSA: SAT Raw Score Summary Statistics for Total Group and Subgroups

Form 1 May 2014		*Mathematics*			*Critical Reading*			*Writing*		
		N	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*
Gender	Male	5,976	22.8	13.7	5,976	26.4	16.0	5,976	20.0	11.0
	Female	5,813	21.2	12.5	5,813	27.9	14.7	5,813	21.8	10.6
	All	11,789	22.0	13.1	11,789	27.2	15.4	11,789	20.9	10.8
Students with disabilities	Yes	1,158	8.5	9.5	1,158	12.5	12.9	1,158	10.4	8.5
	No	10,631	23.5	12.6	10,631	28.8	14.8	10,631	22.1	10.4
	All	11,789	22.0	13.1	11,789	27.2	15.4	11,789	20.9	10.8
Economically disadvantaged	Yes	4,040	16.6	11.3	4,040	21.4	14.0	4,040	16.7	9.6
	No	7,749	24.8	13.1	7,749	30.2	15.2	7,749	23.1	10.8
	All	11,789	22.0	13.1	11,789	27.2	15.4	11,789	20.9	10.8
Limited English Proficiency	Currently receiving LEP	198	10.0	10.6	198	8.9	9.4	198	9.6	7.5
	Formerly received LEP	94	16.6	10.4	94	18.7	9.1	94	15.2	7.1
	No LEP	11,497	22.3	13.1	11,497	27.5	15.3	11,497	21.1	10.8

SD = standard deviation

Appendix K contains scaled-score distribution graphs showing the relative and cumulative percentages of students at each scaled score. The total number (N) of students tested is also given, from which the number of students assigned each scaled score can be derived. Appendix K also shows, in Table K-1, achievement-level distributions for each of the last three administrations.

Table 8-4 below shows the MHSA scaled-score ranges that correspond to each achievement level.

Table 8-4. 2013–14 MHSA:SAT Range of Scores for Each Achievement Level

Content Area	*Substantially Below Proficient*	*Partially Proficient*	*Proficient*	*Proficient with Distinction*
Mathematics	1100–1132	1134–1140	1142–1160	1162–1180
Critical reading	1100–1128	1130–1140	1142–1160	1162–1180
Writing	1100–1128	1130–1140	1142–1160	1162–1180

8.4 Classification Accuracy and Consistency of MHSA: SAT Cut Scores

While related to reliability, the accuracy and consistency of classifying students into achievement categories are even more important statistics in a standards-based reporting framework (Livingston and Lewis, 1995). After the achievement levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the MHSA, students are classified into one of four achievement levels: Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction. This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2013–14 MHSA because it is easily adaptable to all types of testing formats, including mixed-format tests.

The accuracy and consistency estimates reported below make use of “true scores” in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to categorize students into their “true” classifications.

For the 2013–14 MHSA, after various technical adjustments (described in Livingston and Lewis, 1995), a four-by-four contingency table of accuracy was created where cell $[i, j]$ represented the estimated proportion of students whose true score fell into classification i (where $i = 1$ to 4) and observed score into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

To calculate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments per Livingston and Lewis (1995), a new four-by-four contingency table was created and populated by the proportion of students who would be categorized into each combination of classifications according to the two (hypothetical) parallel test forms. Cell $[i, j]$ of this table represented the estimated proportion of students whose observed score on the first form would fall into classification i (where $i = 1$ to 4) and whose observed score on the second form would fall into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students categorized by the two forms into exactly the same classification) signified overall consistency.

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{(\text{Observed agreement})-(\text{Chance agreement})}{1-(\text{Chance agreement})} = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1-\sum_i C_{i.}C_{.i}}$$

where
$C_{i.}$ is the proportion of students whose observed achievement level would be Level i (where $i = 1$–4) on the first hypothetical parallel form of the test;
$C_{.i}$ is the proportion of students whose observed achievement level would be Level i (where $i = 1$–4) on the second hypothetical parallel form of the test;
C_{ii} is the proportion of students whose observed achievement level would be Level i (where $i = 1$–4) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than other consistency estimates.

8.4.1 Accuracy and Consistency

Results of the accuracy and consistency analyses described above are provided in Table 8-5. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, for Mathematics, the conditional accuracy value is 0.85 for Substantially Below Proficient. This figure indicates that among the students whose true scores placed them in this classification, 85% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.80 indicates that 80% of students with observed scores in the Substantially Below Proficient level would be expected to score in this classification again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, in testing done for NCLB accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, the accuracy of the Partially Proficient/Proficient threshold is of greatest interest. For the 2013–14 MHSA, Table 8-6 provides accuracy and consistency estimates at each cutpoint as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

The above indices are derived from Livingston and Lewis's (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: (1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and (2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

Note that, as with other methods of evaluating reliability, DAC statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Tables 8-5 and 8-6 should be interpreted with caution.

Table 8-5. 2013-14 MHSA: SAT Summary of Decision Accuracy (and Consistency) Results by Subject—Conditional on Cutpoint

Subject	*Grade*	*Substantially Below Proficient / Partially Proficient*			*Partially Proficient / Proficient*			*Proficient / Proficient with Distinction*		
		Accuracy (consistency)	*False*		*Accuracy (consistency)*	*False*		*Accuracy (consistency)*	*False*	
			Positive	*Negative*		*Positive*	*Negative*		*Positive*	*Negative*
Mathematics	11	0.93 (0.90)	0.04	0.03	0.92 (0.88)	0.05	0.04	0.98 (0.97)	0.02	0.01
Reading	11	0.93 (0.91)	0.03	0.03	0.92 (0.89)	0.05	0.03	0.96 (0.95)	0.03	0.01
Science	11	0.91 (0.87)	0.05	0.05	0.89 (0.85)	0.06	0.05	0.98 (0.97)	0.02	0.01
Writing	11	0.92 (0.89)	0.04	0.04	0.90 (0.86)	0.06	0.04	0.96 (0.95)	0.03	0.01

Table 8-6. 2013-14 MHSA: SAT Summary of Decision Accuracy (and Consistency) Results by Subject—Overall and Conditional on Performance Level

Subject	*Grade*	*Overall*	*Kappa*	*Conditional on level*			
				Substantially Below Proficient	*Partially Proficient*	*Proficient*	*Proficient with Distinction*
Mathematics	11	0.82 (0.76)	0.64	0.86 (0.81)	0.69 (0.59)	0.88 (0.83)	0.85 (0.69)
Reading	11	0.81 (0.74)	0.64	0.86 (0.81)	0.73 (0.63)	0.84 (0.78)	0.87 (0.74)
Science	11	0.79 (0.72)	0.56	0.84 (0.78)	0.52 (0.41)	0.87 (0.81)	0.80 (0.58)
Writing	11	0.78 (0.70)	0.57	0.82 (0.76)	0.70 (0.61)	0.82 (0.75)	0.84 (0.66)

8.5 COMPLETION RATES

Completion rate refers to the extent to which the test takers are able to complete each section of the test in the time allotted. Because there is no generally accepted index of acceptable or adequate completion rates, several criteria are reported. Each is arbitrary and by itself should not be too strictly applied. However, taken together, the criteria can be useful. When considering these criteria, the relative ability of the group, as defined by the analysis sample scaled-score mean and median, needs to be taken into account.

One statistic reported is the percentage of the analysis sample reaching the items at the end of each test section. These results may be confounded with item difficulty because one or two very difficult items at the end of the test section may make it appear more speeded than it really is. This case would be shown by a sharp decrease in the number of test takers completing the last few items, rather than a gradual tapering off.

Additional completion rate data are based on the items that are not reached. Information presented in Table 8-7 includes the percentage of the group who completed each section (answered the last item in the section), the percentage of the group who completed 75% of the section (answered one or more items that were at least three-quarters of the way through the section), and the number of items that were reached by 80% of the group. The ratio of the variance of the number of items not reached to the variance of the formula scores (given as "NR variance/score variance") is presented in the table as another index of completion rate. The total number of items in each section and the mean and standard deviation of the number of items not reached are also given in the table.

As a rule of thumb, a test is usually regarded as essentially unspeeded if at least 80% of the test takers reach the last question and if virtually everyone reaches at least three-quarters of the items. Swineford (1974) determined that a variance index less than 0.15 may be taken to indicate an unspeeded test, while an index greater than 0.25 usually means that the test is clearly speeded. Values between 0.16 and 0.25 generally indicate a moderately speeded test. However, these are only arbitrary indices, and judgments of appropriateness of timing should be made in the context of additional data. For example, lack of motivation among the test takers may make sections appear more speeded.

Table 8-7 provides the speededness data for the state of Maine. The May 2014 critical reading portion is unspeeded with 80% of examinees reaching the last item for two of the three sections in May and with all of the 3 variance indices below 0.15. The Critical Reading section 3 was only slightly speeded with 80% of examinees reaching 18 of the 19 items and a variance index of 0.11. The June 2014 critical reading portion is also essentially unspeeded with the exception of section 2 which is slightly speeded with 80% of examinees reaching 23 of the 24 items and a variance index of 0.04. The mathematics sections are slightly speeded with less than 80% of students reaching the last item on all 6 sections. However, all variance indices are at or below 0.15 with the exception of Mathematics section 2 in June with an index of 0.20. The writing portion is unspeeded for both May and June 2014 though less than 80% of the students were able to reach the last item in Writing section 1 of both administrations. The variance indices for Writing are all below 0.15. Completion rate data for the national SAT population are provided in Appendix D, Table D-8.

Table 8-7. 2013–14 MHSA: Maine Completion Rate Statistics for Sections of the College Board SAT

Form	*1*	*2*	*1*	*2*	*1*	*2*
Administration	05/14	06/14	05/14	06/14	05/14	06/14
*Sample size**	10,411	188	10,411	188	10,411	188
	Critical Reading 1		*Mathematics 1*		*Writing 1*	
% completing section	83.7	84.0	57.7	57.4	73.4	68.6
% completing 75%	99.8	98.9	95.6	98.9	100.0	100.0
Number of items reached by 80%	23	24	18	19	34	34
Mean not reached	0.3	0.6	1.2	0.7	0.7	0.7
SD not reached	0.9	1.9	1.9	1.2	1.3	1.2
NR variance/score variance	0.02	0.13	0.14	0.07	0.03	0.02
Number of items	23	24	20	20	35	35
	Critical Reading 2		*Mathematics 2*		*Writing 2*	
% completing section	83.5	78.2	39.4	41.0	87.8	94.7
% completing 75%	97.0	100	94.0	89.4	98.7	100
Number of items reached by 80%	25	23	15	16	14	14
Mean not reached	0.7	0.6	1.4	1.4	0.2	0.1
SD not reached	1.9	1.2	1.8	2.0	0.8	0.4
NR variance/score variance	0.10	0.04	0.15	0.20	0.05	0.01
Number of items	25	24	18	18	14	14
	Critical Reading 3		*Mathematics 3*			
% completing section	73.2	84.0	70.4	64.4		
% completing 75%	96.7	96.8	97.6	95.7		
Number of items reached by 80%	18	19	15	14		
Mean not reached	0.7	0.5	0.6	0.9		
SD not reached	1.5	1.6	1.3	1.7		
NR variance/score variance	0.11	0.12	0.09	0.15		
Number of items	19	19	16	16		

* The sample size is the final sample of Maine NCLB students taking the test and answering at least one question in each respective section of the test.

SD = standard deviation; NR = number of items not reached

8.6 Item Statistics

8.6.1 Item Difficulty: Equated Delta

The simplest measure of item difficulty for a given group of test takers is the p-value—the proportion of test takers who attempted to answer the item correctly compared to those who attempted to answer the item. For the SAT, p-values are converted onto a standard scale called the delta index.

$$Delta = 13 + 4z$$

where
z is computed based on item difficulty, p.

First, $(1 - p)$ is converted to a normalized z-score and then linearly transformed to a scale with a mean of 13 and a standard deviation of 4. Deltas are inversely related to p-values; that is, the lower the p-value, the higher the delta, and the more difficult the item.

The conversion of p-values provides raw delta values that reflect the difficulty of the items for the particular test takers from a particular administration. This measure of item difficulty then must be adjusted to correct for differences in the abilities of different test-taking populations. Delta equating is a statistical procedure used to convert raw delta values to equated delta values. This procedure involves administering some old items with known equated delta values along with new items. Each old item now has two difficulty measures: the observed delta that reflects the difficulty of the item for the current group of test takers and the equated delta that is an estimate of how difficult the items would have been for the initial reference group. The linear relationship between the pairs of observed and equated deltas on the old items is used to determine the scaled values for each of the new items. Delta equating is essential because the groups taking a particular test may differ substantially in ability from one administration to another. Through delta equating, item difficulties can be compared directly.

As described in Chapter 2, new forms of the SAT are built to detailed content and statistical specifications. Each item in the new form has already been administered and has an associated difficulty estimate (equated delta). SAT statistical specifications set target means and standard deviations of the equated deltas for mathematics, critical reading, and writing. In addition, each measure has a specific requirement for the particular number of items at each delta level across the range of the delta scale. For each measure, the delta distribution is a unimodal distribution with more middle difficulty items and fewer very easy or very difficult items. The target mean delta is 11.4 (standard deviation of 2.4) for critical reading. The means and standard deviations of the deltas for critical reading in May and June 2014 were 11.5 (2.3) and 11.6(2.2), respectively, which are in range to be considered as having met the specifications for the section. For mathematics and writing, the mean deltas for the two forms administered in May and June 2014 are also very close to the specifications and within an acceptable range of variation, though the mean for the math components are slightly below the target on the June form. Table 8-8 summarizes the mean equated delta and standard deviation for each content area by form for students testing on the MHSA in Maine only.

Table 8-8. 2013–14 MHSA: Maine Summary Statistics of Equated Deltas (Δ) for Mathematics, Critical Reading, and Writing Sections of the College Board SAT*

			Form 1	*Form 2*
			May 2014	June 2014
			11,789	205
Content Area		*Specified Equated Delta*	*Equated Delta*	*Equated Delta*
Total Critical Reading	N	67	67	66
	Mean	11.4	11.5	11.6
	S.D.	2.4	2.3	2.2
Mathematics – Multiple Choice	N	44	44	44
	Mean	12.2	12.2	12.1
	S.D.	3.2	3.1	3.0
Mathematics SPR	N	10	10	10
	Mean	13.6-14.2	13.8	13.4
	S.D.	3.0	2.7	2.7
Total Writing	N	49	49	49
	Mean	10.1	10.1	10.1
	S.D.	2.5	2.4	2.5

* Estimates are based on students who took the MHSA SAT component and answered at least one item in each section.

MC = multiple-choice; SPR = student-produced response; SD = standard deviation

8.6.2 Item Discriminating Power: Biserial Correlation

Another important characteristic of an item is item discrimination. Each item in a test should be able to distinguish higher-ability test takers from lower-ability test takers with respect to the construct being tested. An item is considered discriminating if proportionately more test takers who are high in the ability being measured answer the item correctly than do test takers low in the ability being measured. The total score is generally used as the criterion for judging levels of ability on the construct being tested. Item difficulty can constrain item discrimination power, in that if most or very few examinees are responding correctly to an item, the discrimination is restricted.

A number of indices are used in assessing the discriminating power of an item. The index currently used on the SAT is the biserial correlation coefficient, which measures the strength of the relationship (correlation) between examinees' performance on a single item and the formula score, excluding the item being analyzed. A very low or negative correlation indicates that the item does not add any precision to the measurement of the test as a whole.

During assembly of new forms, there are specifications concerning discrimination. The specified mean biserial for both critical reading and writing is 0.49 to 0.53. For mathematics, the specified mean biseral is 0.53 to 0.57 on the multiple-choice items and 0.60 to 0.70 on the student-produced-response items. Table 8-9 presents the biserial coefficients for the May and June 2014 forms of the SAT for students taking the

MHSA in Maine only. All values are within the specified range on the May 2014 form with the exception of the Writing section which is slightly below the range. The June 2014 form is slightly outside the specified range but still within an acceptable degree of variance.

Table 8-9. 2013–14 MHSA: Maine Summary Statistics for Biserial Coefficients[1] for Mathematics, Critical Reading, and Writing Sections of the College Board SAT

			Form 1	*Form 2*
			May 2014	June 2014
			11,789	205
Content Area		*Specified Ranges*		
Total Critical Reading	N	0.49-0.53	67	66
	Not Comp.[2]			1
	Mean		0.50	0.47
	S.D.		0.12	0.15
Mathematics – Multiple Choice	N	0.53-0.57	44	44
	Not Comp.[2]			
	Mean		0.56	0.51
	S.D.		0.12	0.15
Mathematics SPR	N	0.60-0.70	10	10
	Not Comp.[2]			
	Mean		0.67	0.73
	S.D.		0.12	0.12
Total Writing	N	0.49-0.53	49	49
	Not Comp.[2]			
	Mean		0.47	0.48
	S.D.		0.11	0.11

[1] Estimates are based on students who took the MHSA SAT component and answered at least one item in each section.

[2] An *r*-biserial is not calculated when the percentage correct is greater than 95 or less than 5, or when dropout exceeds 50%

MC = multiple-choice; SPR = student-produced response; SD = standard deviation

8.7 Differential Item Functioning

Measures of differential item functioning (DIF) are used to help ensure test and item fairness. DIF indicates "a difference in item performance between two comparable groups of examinees; that is, the groups that are matched with respect to the construct being measured by the test" (Dorans and Holland, 1993, p. 35). Theoretically, if test takers from two different groups have the same ability level, they should have the same probability of getting an item correct. The two groups are referred to as the focal group and the reference group, where the focal group is the focus of analysis and the reference group is the basis for comparison.

Currently, the SAT uses the Mantel-Haenszel (MH) approach (Holland and Thayer, 1988) for DIF detection (D-DIF). On the basis of the MH D-DIF statistic, which can be interpreted as a difference in deltas, items are classified into the following categories based on specific criteria:

- Category A—Negligible DIF: Items are classified in this category for a particular combination of reference and focal groups if either MH D-DIF is not statistically different from 0 or if the magnitude of the MH D-DIF value is less than 1.0 delta unit in absolute value.
- Category B—Intermediate DIF: This category is composed of items that are not classified as A or C
- Category C—Large DIF: Items are classified as C if MH D-DIF both exceeds 1.5 in absolute value and is statistically significantly larger than 1.0 in absolute value.

A minus sign (e.g., B- or C-) indicates that the item tended to favor the reference group (male or White), while a plus sign (e.g., B+ or C+) indicates the item tended to favor the focal group (female or non-White).

The current practice for the SAT is to run DIF for selected ethnicities, with Whites as the reference group. Separate DIF analyses are performed with African Americans, Hispanics, Asian Americans, and Native Americans as the focal groups. In Maine, the population is not as diverse as that found nationally; therefore, subgroup sample sizes permitted only analyses for the African American versus White ethnicity comparison. DIF analyses are also performed with males as the reference group and females as the focal group. The DIF analyses completed using all students who took the May and June 2014 SAT test forms for the national population are listed in Tables D-11 and D-12 of Appendix D. Table 8-10 represents DIF analyses for Form 1 of the SAT using only students from Maine. DIF analyses for the June administration, Form 2, were not conducted due to insufficient sample size.

For the analysis using only Maine students, fewer students were available. The low number of students had two immediate impacts upon the analysis. First, comparisons across all groups were not possible. A standard minimum applied when completing DIF analysis is that 200 or more students must exist in each group being analyzed. Using a sample of students fewer than 200 would yield unreliable results. While the sample for the African American students exceeds the criteria of 200 students, some caution should be used in the interpretation of these results as well. A potential second impact of the small sample sizes is that more items may have been classified with C-DIF. In May 2014, no items were classified with C-DIF.

Table 8-10. 2013–14 MHSA: Maine Differential Item Functioning (DIF) Summary Form: 1 Administration: 5/14

Category of Maximum Absolute DIF Value for All Comparisons				*Female* N = 5,813	*African American* N = 349
				Male N = 5,976	*White* N = 10,896
Content Area	*Category*	*Number*	*% of Items*	*Number of Items by DIF Category*	
Total critical reading	+C	0	0.0	0	0
	+B	0	0.0	0	0
	A	62	92.5	63	66
	-B	5	7.5	4	1
	-C	0	0.0	0	0
	Total	67	100	67	67
Total mathematics	+C	0	0.0	0	0
	+B	3	5.6	2	2
	A	47	87.0	48	51
	-B	4	7.4	4	1
	-C	0	0.0	0	0
	Total	54	100	54	54
Total writing	+C	0	0.0	0	0
	+B	4	8.2	0	4
	A	42	85.7	48	43
	-B	3	6.1	1	2
	-C	0	0.0	0	0
	Total	49	100	49	49

8.8 Summary

The scores reported for SAT test takers must be accurate and comparable regardless of which form is administered or at which administration the student takes the examination. The intention of this chapter was to describe the intense scrutiny that each item, form, and reported score must undergo. The care and thought required in establishing a new scale, such as the new writing section, and in maintaining the meaning of established scales, such as the mathematics and critical reading sections, were also described. The information in this chapter should help the reader to understand the psychometric rigor required to ensure that the interpretations of the score results are valid and fair. In addition, the statistical results that were reported concerning items and forms provide assurance that the test scores are reliable. For information on interpreting SAT scores, visit http://www.collegeboard.com/prod_downloads/sat/sat-educators-handbook.pdf and see pages 22–26.

CHAPTER 9 THE MHSA SCIENCE COMPONENT

9.1 CLASSICAL ITEM ANALYSIS

As noted in Brown (1983), “A test is only as good as the items it contains.” A complete evaluation of a test’s quality must include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (AERA et al., 1999) and *Code of Fair Testing Practices in Education* (2004) include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. Items should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. In addition, items must not unfairly disadvantage students in particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that MHSA science items meet these standards. Qualitative analyses are described in earlier chapters of this report; this chapter focuses on quantitative evaluations. Statistical evaluations are presented in four parts: 1) difficulty indices, 2) item-test correlations, 3) differential item functioning (DIF) statistics, and 4) dimensionality analyses. The item analyses presented here are based on the statewide administration of the MHSA science test in spring 2014.

Note that, to facilitate interpretability of the calculated statistics, formula scoring of multiple-choice items was not implemented for purposes of calculating classical difficulty and discrimination indices or DIF statistics.

9.1.1 Classical Difficulty and Discrimination Indices

All multiple-choice and constructed-response items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. For purposes of calculating classical item statistics, the multiple-choice items were scored dichotomously (i.e., without formula scoring); therefore, for these items, the difficulty index is simply the proportion of students who correctly answered the item. Constructed-response items are scored polytomously, meaning that a student can achieve a score of 0, 1, 2, 3, or 4. By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale, ranging from 0.0 to 1.0 regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an easiness index, because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student abilities, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students provide little information about differences in student abilities but may indicate knowledge or skills that have not yet been mastered by most students. In general, to provide the best measurement, difficulty indices should range from near-chance performance (0.25 for four-option multiple-choice items or essentially zero for constructed-response items) to 0.90, with the majority of items generally falling between 0.4 and 0.7. However, on a standards-referenced assessment such as the MHSA science test, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students do. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of the item. Within classical test theory, the item-test correlation is referred to as the item's discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For constructed-response items, the item discrimination index used was the Pearson product-moment correlation. For the multiple-choice items, formula scoring was not implemented for purposes of calculating classical item statistics, so the item discrimination index used was the point-biserial correlation. The theoretical range of these statistics is -1.0 to 1.0, with a typical observed range from 0.2 to 0.6.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency.

A summary of the item difficulty and item discrimination statistics is presented in Table 9-1. Note that the statistics are presented for all items as well as by item type (multiple-choice and constructed-response). The mean difficulty and discrimination values shown in the table are within generally acceptable and expected ranges.

Table 9-1. 2013–14 MHSA Science: Summary of Item Difficulty and Discrimination Statistics

Item Type	*Number of Items*	*p-Value*		*Discrimination*	
		Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ALL	44	0.54	0.16	0.34	0.10
MC	40	0.56	0.15	0.32	0.08
CR	4	0.35	0.08	0.52	0.12

MC = multiple-choice; CR = constructed-response

Comparing the difficulty indices of multiple-choice items and constructed-response items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that the difficulty indices for multiple-choice items tend to be higher (indicating that students performed better on these items) than the difficulty indices for constructed-response items. Similarly, discrimination indices for the four-point constructed-response items were larger than those for the dichotomous items because of the greater variability of the former (i.e., the partial credit these items allow) and the tendency for correlation coefficients to be higher given greater variances of the correlates.

In addition to the item difficulty and discrimination summaries presented above, item-level classical statistics and item-level score-point distributions were also calculated. Item-level classical statistics are provided in Appendix E; item difficulty and discrimination values are presented for each item. The item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There were a small number of items with low discrimination indices, but none were negative. While it is not inappropriate to include items with low discrimination values or with very high or very low item difficulty values to ensure that content is appropriately covered, there were very few such cases on the MHSA science test. Item-level score-point distributions are provided for constructed-response science items in Appendix F; for each science item, the percentage of students who received each score point is presented.

9.1.2 Differential Item Functioning

Code of Fair Testing Practices in Education (2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are because of construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines. As part of the effort to identify such problems, MHSA science items were evaluated in terms of DIF statistics.

For the MHSA science test, the standardization DIF procedure (Dorans and Kulick, 1986) was employed to evaluate subgroup differences. The standardization DIF procedure is designed to identify items for which subgroups of interest perform differently, beyond the impact of differences in overall achievement. The DIF procedure calculates the difference between item performance for two groups of students (at a time) matched for achievement on the total test. Specifically, average item performance is calculated for students at every total score. Then an overall average is calculated, weighting the total score distribution so that it is the same for the two groups.

When differential performance between two groups occurs on an item (i.e., a DIF index in the “low” or “high” categories, explained below), it may or may not be indicative of item bias. Course-taking patterns or differences in school curricula can lead to DIF, but for construct-relevant reasons. On the other hand, if subgroup differences in performance could be traced to differential experience (such as geographical living conditions or access to technology), the inclusion of such items should be reconsidered.

Computed DIF indices have a theoretical range from -1.0 to 1.0 for multiple-choice items, and the index is adjusted to the same scale for constructed-response items; here, again, formula scoring was not applied to the items for purposes of calculating DIF statistics. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The preponderance of MHSA science items fell within this range. Dorans and Holland further stated that items with values between -0.10 and -0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked and that items with values outside the -0.10 to 0.10 range (i.e., "high" DIF) are more unusual and thus should be examined very carefully.

For the 2013-14 MHSA, five subgroup comparisons were evaluated for DIF:

- male versus female
- no disability versus disability
- not economically disadvantaged versus economically disadvantaged
- non-limited English proficient (LEP) versus LEP
- white (non-Hispanic) versus black or Asian American

The table in Appendix G presents the number of items classified as either "low" or "high" DIF overall and by group favored.

9.1.3 Dimensionality Analyses

The MHSA science test was designed to measure and report a single score on science achievement using a unidimensional scale from 1100 to 1180. Thus, this test is said to be measuring a single dimension, and the term *unidimensional* is used to describe such a test.

Because the high school science test was constructed with multiple content area subcategories, and their associated knowledge and skills, the potential exists for a large number of secondary dimensions being invoked beyond the primary science dimension that all the items have in common. Generally, the scores on such subtests are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional item response theory (IRT) models that are used for calibrating, linking, scaling, and equating the 2013–14 MHSA science test forms.

The purpose of dimensionality analysis is to investigate whether violations of the assumption of test unidimensionality are statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Findings from dimensionality analyses performed on the 2013–14 MHSA science test are reported below. (Note: Only common items were analyzed since they are used for score reporting.)

Dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, and Gao, 2001) and DETECT (Zhang and Stout, 1999). Nonparametric techniques were preferred for this analysis because such techniques avoid strong parametric modeling assumptions while still adhering to the fundamental principles of IRT. Parametric techniques, such as nonlinear factor analysis, make strong assumptions that are often inappropriate for real data, such as assuming a normal distribution for ability and lower asymptotes of zero for the item characteristic curves.

Both DIMTEST and DETECT use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on expected total score for the rest of the test, and the average conditional covariance is obtained by averaging all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Nonzero conditional covariances are essentially violations of the principle of local independence, and local dependence implies multidimensionality. Thus, nonrandom patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. For exploratory analyses, the data are first randomly divided into a training sample and a cross-validation sample. Then an analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items display local dependence, conditioning on total score on the nonclustered items. For confirmatory analyses, the practitioner selects a group of items suspected to represent a secondary dimension, and the whole sample is used to test whether the conditional covariances of the selected cluster of items display local dependence, conditioning on total score on the nonclustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. For exploratory analyses, as with DIMTEST, the data are first randomly divided into a training sample and a cross-validation sample (if a DIMTEST exploratory analysis has been conducted, one could use the same training and cross-validation samples as were used with DIMTEST, but using new samples is also permissible). The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed; from this sum the between-cluster conditional covariances are subtracted; this difference is divided by the total number of item pairs; and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. For confirmatory analyses, the practitioner selects the clusters, and then the DETECT statistic is calculated in the same way as for exploratory analyses, but using all the data, not just the cross-validation sample. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality); values of 0.2 to 0.4, weak to moderate multidimensionality;

values of 0.4 to 1.0, moderate to strong multidimensionality; and values greater than 1.0, very strong multidimensionality.

DIMTEST and DETECT were applied to the 2013–14 MHSA science test. The data were first split into a training sample and a cross-validation sample. Because the total sample size was over 12,600 student examinees, the training sample and cross-validation sample each had more than 6,300 students.

DIMTEST was then applied to the MHSA science test. Because of the very large sample size of this test, DIMTEST would be sensitive even to quite small violations of unidimensionality; the null hypothesis was rejected with a *p*-value less than 0.00005. The occurrence of statistical rejection of the null hypothesis was not surprising because strict unidimensionality is an idealization that rarely holds exactly for a given dataset. Thus, it was important to use DETECT to estimate the effect size of the violation of local independence found by DIMTEST.

Next, a DETECT analysis was conducted on the MHSA science test. This resulted in a DETECT statistic of 0.19, a value indicative of very weak multidimensionality. Furthermore, the ratio of the DETECT statistic to the maximum possible value of the DETECT statistic was only 0.49, and the percentage of conditional covariance pairs having positive signs for item pairs in the same cluster and negative signs for items coming from different clusters was only 63.9%.

The clusters reported by DETECT were investigated, and they indicated some tendency for the four 4-point open-response (OR4) items to cluster separately from the multiple-choice (MC) items. Specifically, there was one cluster that contained all the OR4 items but only 12 MC items. Thus, in this cluster the OR4 items accounted for about 54% of the points, whereas on the test as a whole the OR4 items only accounted for about 29% of the total points. The OR4 items in this cluster had many positive conditional covariances with each other, but they also had a substantial number of positive conditional covariances with MC items (instead of the almost all negative conditional covariances one would expect if they were a strongly distinct dimension).

Finally, we note that these results are very similar to the results from the analyses of the MHSA science tests for the previous six years (from 2007–08 to 2012–13) In particular, for these past years, rejection of the DIMTEST null hypothesis of unidimensionality occurred every year, with the largest *p*-value being 0.0005. The DETECT effect sizes for the previous six years were 0.23, 0.17, 0.15, 0.13, 0.11, and 0.04, respectively, for 2007–08 through 2012–13, indicating either weak or very weak multidimensionality every year.

Taken together, the DIMTEST and DETECT results for the science test indicate that the test has very small, though detectable, violations of unidimensional local independence, and that these violations are at least partially related to the two item types used on the test. Thus, although these results indicate some definite multidimensionality, it is very weak in magnitude. Therefore, no changes in test design or scoring for the science test seem to be warranted in regard to multidimensionality. In particular, the dimensionality analysis results support the application of unidimensional IRT to the MHSA science test for purposes of calibrating, linking, scaling, and equating. Indeed, the results support using unidimensional IRT to place the MHSA science items onto a single score scale for reporting purposes.

9.2 Item Response Theory Scaling and Equating

The MHSA science test uses a pre-equating model in which items are calibrated using IRT and placed on scale at the time of field testing. These item parameters are then used to assemble test forms that meet content blueprints and psychometric quality criteria. The sections below describe the procedures used to calibrate the MHSA items and to calculate scaled scores and achievement levels used for reporting.

9.2.1 Item Response Theory

As mentioned above, all MHSA science items were calibrated using IRT. IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (θ), and the probability (p) of getting a dichotomous item correct or of getting a particular score on a polytomous item. In IRT, it is assumed that all items are independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and p (Hambleton and van der Linden, 1997; Hambleton and Swaminathan, 1985). The process of determining the specific mathematical relationship between θ and p is called item calibration. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and p. Once the item parameters are known, an estimate of θ for each student can be calculated. This estimate, $\hat{\theta}$, is considered to be an estimate of the student’s true score or a general representation of student performance. It has characteristics that may be preferable to those of raw scores for equating purposes.

Because of the use of formula scoring, we use a polytomous IRT model for all items. The multiple-choice items are scored such that an incorrect response is given a score of -0.33, an omit is given a score of 0, and a correct answer is given a score of 1 (i.e., formula scoring). The student response records are initially coded as 0, 1, or 2, and the integer scoring function is modified from 0, 1, and 2 to -0.33, 0, and 1 after the IRT calibration and equating process is complete. Thus, the response category probability values as estimated during IRT calibration are multiplied by their respective value from the modified scoring function.

In the graded response model (GRM) for polytomous items, an item is scored in $k + 1$ graded categories that can be viewed as a set of k dichotomies. At each point of dichotomization (i.e., at each threshold), a two-parameter model can be used. This implies that a polytomous item with $k + 1$ categories can be characterized by k item category threshold curves (ICTC) of the two-parameter logistic form:

$$P_{ik}^{*}=\frac{\exp\left[Da_i\left(\theta_j-b_i+d_{ik}\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_{ik}\right)\right]}$$

where
i indexes the items,
j indexes students,
k indexes threshold,
a represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

After computing k ICTCs in the GRM, k + 1 item category characteristic curves (ICCCs) are derived by subtracting adjacent ICTCs:

$$P_{ik}\left(1\middle|\theta_j\right)=P_{i(k-1)}^{*}\left(1\middle|\theta_j\right)-P_{ik}^{*}\left(1\middle|\theta_j\right)$$

where
P_{ik} represents the probability that the score on item i falls in category k, and
P_{ik}^{*} represents the probability that the score on item i falls above the threshold k.
($P_{i0}^{*}=1$ and $P_{i(m+1)}^{*}=0$).

The GRM is also commonly expressed as:

$$P_{ik}\left(k\middle|\theta_j,\xi_i\right)=\frac{\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}-\frac{\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]}$$

where
ξ_i represents the set of item parameters for item i.

Finally, the ICC for polytomous items is computed as a weighted sum of ICCCs, where each ICCC is weighted by a score assigned to a corresponding category:

$$P_i\left(1\middle|\theta_j\right)=\sum_{k}^{m+1}w_{ik}P_{ik}\left(1\middle|\theta_j\right)$$

For more information about item calibration and determination, the reader is referred to Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

9.2.2 Item Response Results

The tables in Appendix H give the IRT item parameters of all common items on the 2013–14 MHSA tests. In addition, Appendix I shows graphs of the test characteristic curves (TCCs) and test information functions (TIFs), which are defined below.

TCCs display the expected (average) raw score associated with each θ_j value between -4.0 and 4.0. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in Section 10.1, the expected raw score at a given value of θ_j is

$$E(X|\theta_j) = \sum_{i=1}^{n} P_i(1|\theta_j)$$

where
i indexes the items (and n is the number of items contributing to the raw score),
j indexes students (here, θ_j runs from -4.0 to 4.0), and
$E(X|\theta_j)$ is the expected raw score for a student of ability θ_j.

The expected raw score monotonically increases with θ_j, consistent with the notion that students of high ability tend to earn higher raw scores than do students of low ability. Most TCCs are “S-shaped”: flatter at the ends of the distribution and steeper in the middle.

The TIF displays the amount of statistical information that the test provides at each value of θ_j. Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM). For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, and Rogers, 1991), as follows:

$$SEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

Compared with the tails, TIFs are often higher near the middle of the θ distribution where most students are located and where most items are sensitive by design.

9.2.3 Achievement Standards

MHSA standards to establish science achievement level cut scores were set in May 2009. The standard-setting meeting and results were discussed in the 2009 technical report and standard-setting report provided at that time. The theta-metric cut scores that emerged from the standard-setting meeting will remain fixed throughout the assessment program unless standards are reset for any reason.

9.2.4 Scaled Scores

9.2.4.1 Description of Scale

Because the θ scale used in IRT calibrations is not readily understood by most stakeholders, reporting scales were developed for the MHSA tests. The reporting scale is a simple linear transformation of the underlying θ scale used in the IRT calibrations. Scaled scores range from 1100 to 1180; the Substantially Below Proficient/Partially Proficient cut was set at 1134 and the Partially Proficient/Proficient cut was set at 1142 and the Proficient/Proficient with Distinction cut was set at 1162. (At the student level, scaled scores were reported as even numbers only.)

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement level scores. School- and SAU-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores (i.e., total number of points) on the MHSA tests were translated to scaled scores using the data analytic process known as scaling. Scaling simply converts from one scale to another. In the same way that a given temperature can be expressed on either Fahrenheit or Celsius scales or the same distance can be expressed in either miles or kilometers, student scores on the 2013–14 MHSA tests can be expressed in raw or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' achievement level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores are reported instead of raw scores. First, because multiple-choice items are formula scored, fractional and negative total raw scores are possible, making them undesirable for use in score reporting. In addition, scaled scores make consistent the reporting of results across years. Due to the fact that different sets of items make up each year's test form, raw cut scores may vary slightly from year to year, but the scaled cut scores remain the same. It is this uniformity across scaled scores that facilitates the understanding of student performance.

9.2.4.2 Calculations

The scaled scores are obtained by a simple translation of ability estimates ($\hat{\theta}$) using the linear relationship between threshold values on the θ metric and their equivalent values on the scaled score metric. Students' ability estimates are based on their raw scores and are found by mapping through the TCC. Scaled scores are calculated using the linear equation

$$SS = m\hat{\theta} + b$$

where
m is the slope and
b is the intercept.

The linear transformation is determined by fixing the 1142 and 1162 values. Table 9-2 presents the scaled score cuts (i.e., the minimum scaled score for getting into the next achievement level). It is important to repeat that the values in Table 9-2 do not change from year to year, because the cut scores along the scale do not change unless standards are reset. Also, in a given year it may not be possible to attain a particular scaled score, but the scaled score cuts will remain the same.

Table 9-2. 2013–14 MHSA Science: Science Scaled Score Cuts and Minimum and Maximum Scores

Minimum	*Scaled Score Cuts*			*Maximum*
	SBP/PP	*PP/P*	*P/PWD*	
1100	1134	1142	1162	1180

SBP = Substantially Below Proficient; PP = Partially Proficient;
P = Proficient; PWD = Proficient with Distinction

Table 9-3 shows the cut scores on θ and the slope and intercept terms used to calculate the scaled scores. Note that the values in Table 9-3 will not change unless the standards are reset.

Table 9-3. 2013–14 MHSA Science: Science Cut Scores (on θ Metric), Intercept, and Slope

θ *Cuts*			*Transformation Constants*	
SBP/P	*PP/P*	*P/PWD*	*Slope*	*Intercept*
-0.3318	0.3616	2.3362	10.12863	1138.337

SBP = Substantially Below Proficient; PP = Partially Proficient;
P = Proficient; PWD = Proficient with Distinction

Appendix J contains raw score to scaled score look-up tables for this year and last year. These are the actual tables that were used to determine student scaled scores, error bands, and achievement levels. The SEMs reported in the look-up tables are conditional standard errors of measurement; that is, the SEM is not the same at all score levels. The term *conditional standard error of measurement* (CSEM) indicates the SEM that is associated with a particular score level.

9.2.4.3 Score Distributions

Appendix K contains scaled score distribution graphs showing the relative and cumulative percentages of students at each scaled score. Appendix K also shows, in Table K-1, achievement level distributions. Because standards for the MHSA science assessment were set in 2009, results are shown for the 2011–12 and 2012–13 administrations as well as the 2013–14 administration.

9.3 RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide a dependable assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true ability. For example, a student may misread an item or mistakenly fill in the wrong bubble when he or she knows the answer. Collectively, extraneous factors that impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores or vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test consistently represent students' ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable. (This is referred to as "test-retest reliability.") A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the "remembering items" problem is to give a different, but parallel, test at the second administration. If student scores on each test correlate highly, the test is considered reliable. (This is known as "alternate forms reliability," because an alternate form of the test is used in each administration.) This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address the latter two problems is to split the test in half and then correlate students' scores on the two half-tests; this in effect treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval and with creating and administering two parallel forms of the test are alleviated. This is known as a "split-half estimate of reliability." If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. Another problem with the split-half method of calculating reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, α (alpha), that eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2013–14 MHSA:

$$\alpha \equiv \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma_{(Y_i)}^2}{\sigma_x^2}\right]$$

where
i indexes the item,
n is the total number of items,
$\sigma_{(Y_i)}^2$ represents individual item variance, and
σ_x^2 represents the total test variance.

9.3.1 Reliability and Standard Errors of Measurement

Table 9-4 presents descriptive statistics, Cronbach's α coefficient, and raw score SEMs for the 2013–14 MHSA science assessment.

Table 9-4. 2013–14 MHSA Science: Raw Score Descriptive Statistics, Cronbach's Alpha, and SEMs

Grade	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
		Maximum	*Mean*	*Standard Deviation*		
11	12,760	56	22.77	11.77	0.87	4.24

9.3.2 Subgroup Reliability

The reliability coefficients presented in the previous section were based on the overall population of students who took the 2013–14 MHSA science test. Appendix L presents reliabilities for various subgroups of interest. Subgroup Cronbach's α's were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are calculated only for subgroups with 10 or more students.

For several reasons, the results of this section should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but also on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix L that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Alternatively, α, which is a type of correlation coefficient, may be artificially depressed for subgroups with little variability (Draper and Smith, 1998). Third, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

9.3.3 Subcategory Reliability

Of even more interest are reliabilities for the science reporting subcategories within MHSA, described in Chapter 3. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix L. Once again as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account. The subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between subtests once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

9.3.4 Interrater Consistency

Chapter 7 of this report describes in detail the processes that were implemented to monitor the quality of the hand-scoring of student responses for science constructed-response items. One of these processes was double-blind scoring: Approximately 10% of student responses were randomly selected and scored independently by two different scorers. Results of the double-blind scoring were used during the scoring process to identify scorers that required retraining or other intervention and are presented here as evidence of the reliability of the MHSA science test. A summary of the interrater consistency results is presented in Table 9-5 below. Results in the table are collapsed across the hand-scored items. The table shows the number of score categories, the number of included scores, the percentage of exact agreement, the percentage of adjacent agreement, the correlation between the first two sets of scores, and the percentage of responses that required a third score. This same information is provided at the item level in Appendix M. These interrater consistency statistics are the result of the processes implemented to ensure valid and reliable hand-scoring of items as described in detail in Chapter 7.

Table 9-5. 2013–14 MHSA Science: Summary of Interrater Consistency Statistics Collapsed Across Items

Grade	*Number of*		*Percent*		*Correlation*	*Percent of Third Scores*
	Score Categories	*Included Scores*	*Exact*	*Adjacent*		
11	5	4,874	60.59	33.38	0.77	5.70

9.3.5 Reliability of Achievement Level Categorization

While related to reliability, the accuracy and consistency of classifying students into achievement categories are even more important statistics in a standards-based reporting framework (Livingston and Lewis, 1995). After the achievement levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical decision accuracy and consistency (DAC) of the classifications. For the MHSA science test, students are classified into one of four achievement levels: Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction. This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2013–14 MHSA science test because it is easily adaptable to all types of testing formats, including mixed-format tests.

The accuracy and consistency estimates reported below make use of "true scores" in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to categorize students into their "true" classifications.

For the 2013–14 MHSA science test, after various technical adjustments (described in Livingston and Lewis, 1995), a four-by-four contingency table of accuracy was created, where cell [i,j] represented the estimated proportion of students whose true score fell into classification i (where i = 1–4) and observed score fell into classification j (where j = 1–4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

To calculate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments per Livingston and Lewis (1995), a new four-by-four contingency table was created and populated by the proportion of students who would be categorized into each combination of classifications according to the two (hypothetical) parallel test forms. Cell [i,j] of this table represented the estimated proportion of students whose observed score on the first form would fall into classification i (where i = 1–4) and whose observed score on the second form would fall into classification j (where j = 1–4). The sum of the diagonal entries (i.e., the proportion of students categorized by the two forms into exactly the same classification) signified overall consistency.

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{\text{(Observed agreement)} - \text{(Chance agreement)}}{1 - \text{(Chance agreement)}} = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1 - \sum_i C_{i.}C_{.i}}$$

where

$C_{i.}$ is the proportion of students whose observed achievement level would be Level i (where $i = 1$–4) on the first hypothetical parallel form of the test;

$C_{.i}$ is the proportion of students whose observed achievement level would be Level i (where $i = 1$–4) on the second hypothetical parallel form of the test; and

C_{ii} is the proportion of students whose observed achievement level would be Level i (where $i = 1$–4) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than other consistency estimates.

9.3.5.1 Accuracy and Consistency

The accuracy and consistency analyses described above are provided in Tables 9-6 and 9-7. Table 9-7 includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.85 for Substantially Below Proficient. This figure indicates that among the students whose true scores placed them in this classification, 85% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.78 indicates that 78% of students with observed scores in the Substantially Below Proficient level would be expected to score in this classification again if a second parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, in testing done for No Child Left Behind (NCLB) accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, the accuracy of the Partially Proficient/Proficient threshold is of greatest interest. For the 2013–14 MHSA science test, Table 9-6 provides accuracy and consistency estimates at each cutpoint, as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

The above indices are derived from Livingston and Lewis's (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: 1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results, and 2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

Note that, as with other methods of evaluating reliability, DAC statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Tables 9-6 and 9-7 should be interpreted with caution.

Table 9-6. 2013–14 MHSA Science: Summary of Decision Accuracy (and Consistency) Results by Subject and Grade—Conditional on Cutpoint

Subject	*Grade*	*Substantially Below Proficient / Partially Proficient*			*Partially Proficient / Proficient*			*Proficient / Proficient with Distinction*		
		Accuracy (Consistency)	*False*		*Accuracy (Consistency)*	*False*		*Accuracy (Consistency)*	*False*	
			Positive	*Negative*		*Positive*	*Negative*		*Positive*	*Negative*
Science	11	0.89 (0.85)	0.05	0.06	0.89 (0.84)	0.06	0.06	0.98 (0.98)	0.01	0.00

Table 9-7. 2013–14 MHSA Science: Summary of Decision Accuracy (and Consistency) Results by Subject and Grade—Overall and Conditional on Performance and Level

Subject	*Grade*	*Overall*	*Kappa*	*Conditional on Level*			
				Substantially Below Proficient	*Partially Proficient*	*Proficient*	*Proficient with Distinction*
Science	11	0.77 (0.70)	0.54	0.85 (0.78)	0.50 (0.40)	0.85 (0.79)	0.76 (0.52)

CHAPTER 10 THE MHSA

All students who participate in the MHSA receive score reports that contain Maine-specific scores on the SAT and science tests. Those students who take the SAT under college-reportable conditions (i.e., without Maine purposes only [MPO] accommodations) also receive SAT score reports directly from the College Board.

10.1 PRIMARY REPORTS

The primary reports for the 2013–14 MHSA are listed below:

- individual student report for parents/guardians
- student results label
- interactive reporting
- school report
- school administrative unit (SAU) report

All reports were distributed to schools and SAUs via a secure Web site hosted by Measured Progress. In addition, printed copies of the student reports were produced for distribution to parents and guardians by schools. Printed student labels were also produced for use by schools. Each of these reports is described in the following subsections, and sample reports are provided in Appendix N.

10.2 INDIVIDUAL STUDENT REPORT FOR PARENTS/GUARDIANS

The front side of the single-page student report includes a letter from the commissioner of education and the MDOE, a description of the achievement levels, and a graph showing state summary results. The back side provides a complete picture of an individual student's performance on the MHSA, divided into two sections. The first section gives the student's overall performance for each content area. The student's scaled scores and achievement levels are shown, both in a table and graphically. The graph shows the range of possible scaled scores, divided up into the four achievement levels. This section also displays the standard error of measurement (SEM) bar for each content area.

The second section of the student report displays the student's achievement level by content area relative to the percentage of students at each achievement level for the school, SAU, and state. For science only, student-level data is displayed by content standard cluster as the number of points attained.

10.3 Student Results Label

To aid schools in keeping track of student scores, schools were supplied with student score information on individual labels that they could affix to school files, if desired.

10.4 Interactive Reporting

There are four interactive reports that were available: item analysis report, achievement level summary, released items summary data, and longitudinal data report. Each of these interactive reports is described in the following sections. Sample interactive reports are provided in Appendix O. To access these four interactive reports, the user clicked the interactive tab on the home page of the system and selected the report desired from the drop-down menu. Next, the user applied basic filtering options, such as the name of the SAU or school and the grade-level test, to open the specific report. At this point, the user had the option of printing the report for the entire grade level or applying advanced filtering options to select a subgroup of students to analyze. Advanced filtering options include gender, ethnicity, limited English proficient (LEP), IEP, and SES. All interactive reports, with the exception of the longitudinal data report, allowed the user to provide a custom title for the report.

10.4.1 Item Analysis Report

The item analysis report provides a roster of all students in a school and provides performance on the items that are released to the public. The student names and identification numbers are listed as row headers down the left side of the report.

For each student, multiple-choice items are marked with either a plus sign (+), indicating that the student chose the correct multiple-choice response, or a letter (from A to D), indicating the incorrect response chosen by the student. For constructed-response items, the number of points earned is shown. All responses to released items are shown in the report, regardless of the student’s participation status. The columns on the right side of the report show the total test results, broken into several categories. Content Strand Points Earned columns show points earned by the student in each content area subcategory relative to total possible points. A Total Points Earned column is a summary of all points earned and total possible points in the content area. The last two columns show the student’s scaled score and achievement level. Students reported as Not Tested are given a code in the achievement level column to indicate the reason the student did not test. It is important to note that not all items used to compute student scores are included in this report; only released items are included. At the bottom of the report, the average percentage correct for each multiple-choice item and average scores for the short-answer and constructed-response items are shown for the school, SAU, and state. When advanced filtering criteria are applied by the user, the School and SAU Percent Correct/Average Score rows at the bottom of the report are blanked out and only the Group row and the State row for the group selected will contain data. This report can be saved, printed, or exported as a PDF, XLS, or CSV file.

The item analysis roster is confidential and should be kept secure within the school and SAU. FERPA requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

10.4.2 Achievement Level Summary

The achievement level summary provides a visual display of the percentages of students in each achievement level for a selected grade. The four achievement levels are represented by various colors in a pie chart. A separate table is also included below the chart that shows the number and percentage of students in each achievement level. This report can be saved, printed, or exported as a PDF or JPG file.

10.4.3 Released Items Summary Data

The released items summary data report is a school-level report that provides a summary of student responses to the released items for a selected grade. The report is divided into two sections by item type (multiple-choice and constructed-response). For multiple-choice items, the total number/percentage of students who answered the item correctly and the number of students who chose each incorrect option or provided an invalid response are reported. An invalid response on a multiple-choice item is defined as "the item was left blank" or "the student selected more than one option for the item." For constructed-response items, point value and average score for the item are reported. Users are also able to view the actual released items within this report. If a user clicks on a particular magnifying glass icon next to a released item number, a pop-up box will open, displaying the released item.

10.4.4 Longitudinal Data Report

The longitudinal data report is a confidential student-level report that provides individual student performance data for multiple test administrations. The state-assigned student identification number is used to link students across test administrations. Student performance on future test administrations will be included on this report over time. This report can be saved, printed, or exported as a PDF file for a single student or for all students within a group.

10.5 School and SAU Reports

Prior to the release of the school and SAU reports to the secure Web site, each SAU office and school received a notification containing a username and password allowing access to these reports. The school and SAU reports consist of three parts: the first part gives an overall summary of scores, the second provides a summary of student participation, and the third includes a report for each content area with scores by reporting subgroups.

The summary of scores includes a table that is designed to show, for each content area, the average scaled score for the school, SAU, and state for each of the last three years, as well as a cumulative average across the three years. In addition, a bar graph for each content area shows the percentage of students in each achievement level at the school, SAU, and state levels. For the SAU version of this report, the school information is blank.

The summary of student participation gives the number and percentage of students who participated at the school, SAU, and state levels for each content area. These numbers are provided for the overall group of students and are broken down by the following categories:

- ethnic group
- identified disability
- LEP status
- socioeconomic status
- migrant status

These numbers are also provided for the overall groups of students, as well as by the following modes:

- students who took the assessment without accommodations
- students who took the assessment with accommodations
- students who took an alternate assessment
- approved nonparticipation in reading for first-year LEP students
- approved nonparticipation for special considerations
- nonparticipation for other reasons

For all three participation modes, data were captured for whether the student had an identified disability or LEP. Again, for the SAU version of this report, the school information is blank.

For each content area, there is a two-page report showing results in more detail. The first page gives a definition of each of the achievement levels along with a table showing the number and percentage of students in the school, SAU, and state who scored at each level. The second page of the content area report breaks the results down by a number of different reporting categories: gender, ethnicity, LEP status, identified disability, socioeconomic status, migrant status, Title 1 program, and 504 plan. This information is provided for the school, SAU, and the state on the school-level report and for the SAU and the state on the SAU-level report. To protect student confidentiality, results are displayed on this page only for groups with 10 or more students.

For each reporting category, the following information is given at the school or SAU level and at the state level:

- the number of students in that category

- the average scaled score for that category
- the percentage of students in the response category who exceeded, met, partially met, or did not meet the standard

10.6 Decision Rules

To ensure that reported results for the 2013–14 MHSA were accurate relative to collected data and other pertinent information, a document that delineates analysis and reporting rules was created. These decision rules were observed in the analyses of MHSA test data and in reporting the assessment results. Moreover, these rules are the main reference for quality-assurance checks.

The decision rules document used for reporting results of the May 2014 administration of the MHSA can be found in Appendix P.

The first set of rules pertains to general issues in reporting scores. Each issue is described and pertinent variables are identified. The actual rules applied are described by the way they impact analyses and aggregations and by their specific impact on each of the reports. The general rules are further grouped into issues pertaining to test items, school type, student exclusions, and number of students for aggregations.

The second set of rules pertains to reporting student participation. These rules describe which students were counted and reported for each subgroup in the student participation report.

10.7 Quality Assurance

Quality assurance measures are embedded throughout the entire process of analysis and reporting. The data processor, data analyst, and psychometrician working on the MHSA implement quality control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within Psychometrics and Research, the sending function verifies that the data are accurate prior to handoff. Additionally, when a function receives a dataset, the first step is to verify the data for accuracy.

Another type of quality assurance measure is parallel processing. Students' scaled scores for each content area are assigned by the psychometrician through a process of equating and scaling. The scaled scores are also computed by the data analyst to verify that scaled scores and corresponding achievement levels are assigned accurately. Respective scaled scores and achievement levels are compared across all students for 100% agreement. Different exclusions assigned to students that determine whether each student receives scaled scores and/or is included in different levels of aggregation are also parallel processed. Using the decision rules document, two data analysts independently write a computer program that assigns students' exclusions. For each content area, the exclusions assigned by each data analyst are compared across all students. Only when 100% agreement is achieved can the rest of data analysis be completed.

The third aspect of quality control involves the procedures implemented by the quality assurance group to check the veracity and accuracy of reported data. Using a sample of schools and SAUs, the quality assurance group verifies that the reported information is correct. The step is conducted in two parts: 1) verify that the computed information was obtained correctly through the appropriate application of different decision rules, and 2) verify that the correct data points populate each cell in the MHSA reports. The selection of sample schools and SAUs for this purpose is very specific and can affect the success of the quality control efforts. There are three sets of samples selected that may not be mutually exclusive. The first set includes those that satisfy the following criteria:

- one-school SAU
- two-school SAU
- multischool SAU

If reporting includes class-level reports, then the set also includes the following:

- multiclass school, multischool SAU
- one-class school, one-school SAU
- multiclass school, one-school SAU
- one-class school, multischool SAU
- private school
- special school (e.g., the "Big 11")
- small school that receives no school report
- small SAU that receives no SAU report
- SAU that receives a report, but all schools are too small to receive a school report
- school with excluded (not tested) students
- school with homeschooled students

The second set of samples includes SAUs or schools that have unique reporting situations as indicated by decision rules. This set is necessary to check that each rule is applied correctly. The third set includes SAUs and schools identified by the MDOE for its review and approval before reports are produced for distribution.

The quality assurance group uses a checklist to implement its procedures. Once the checklist is completed, sample reports are circulated for psychometric checks and program management review. The appropriate sample reports are then sent to the MDOE for review and signoff. Once the MDOE gives the approval to proceed, the reports are posted to Measured Progress's Web site for school and SAU access. Prior to public release, schools and SAUs have a two-week review period in which to examine their results and, if necessary, to report any data issues.

CHAPTER 11 VALIDITY RESEARCH ON THE MHSA SAT COMPONENT

This chapter seeks to bring together a wide range of validity evidence regarding the MHSA SAT Component in a logical and systematic manner. It is guided by the concept of validity articulated in *Standards for Educational and Psychological Testing* (AERA et al., 1999), which provides the following definition: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests." Further, "The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (AERA et al., 1999, p. 9). The purpose of this chapter is to provide some of the more recent evidence supporting the interpretation of SAT scores. Some evidence relates to test content, some to the processes used in responding to the test, some to the internal structure of the test, and still more to the relationship of test scores to other variables, especially criteria such as performance in particular content areas or college grades.

11.1 Construct Validity

The SAT is described variously as a measure of the skills you have learned in and outside of the classroom and how well you can apply that knowledge"[5] or as a test of "the subject matter learned by students in high school and how well they apply that knowledge – the critical thinking skills necessary to succeed in college"[6] What is the nature of the reasoning or critical thinking that is measured by the SAT? Powers and Dwyer (2003) seek to delineate a construct of reasoning, broadly conceived. They point out that "a construct provides a target for a particular assessment; it is not synonymous with the test itself" (p. 1). They identify several definitions of reasoning that have been used by educators, philosophers, and psychologists and note that more recent conceptions of reasoning have emphasized the importance of domain-specific reasoning, i.e., reasoning that is knowledge based. Similarly, a considerable range of definitions for thinking or critical thinking exists. They conclude that there is no single construct of reasoning but that any of the several formulations may be useful and informative depending on the context and purpose.

Powers and Dwyer (2003) argue for the importance of reasoning in academic contexts, such as performance in college. "But of the many things that matter, two of the most important, we believe, are: (a) academic knowledge and skill in the domain of study, and (b) the ability to reason well in the symbol systems used to communicate new knowledge. Reasoning tests correlate with academic success because reasoning abilities are very often required in school learning, whether for understanding a story, inferring the meaning of an unfamiliar word, detecting patterns and regularities in information, going beyond the information given

[5] From Getting Ready for the SAT by the College Board, 2011, p.3.
[6] From About the SAT, retrieved from http://professionals.collegeboard.com/testing/sat-reasoning/about on October 2, 2011.

to form more general rules or principles, or applying mathematical concepts to solve a problem. In these ways and in hundreds of others, successful learning requires reasoning strategies" (p. 12).

This argument seems particularly apropos to the stated purpose of the SAT as a tool in counseling and admissions decisions regarding future learning opportunities. Out of the many possible facets of reasoning, the College Board has chosen to assess three dimensions that are closely related to academic performance: verbal reasoning in the form of critical reading, quantitative reasoning using a defined domain of academic knowledge, and writing—the productive use of a symbol system to communicate one's ability to present and support a point of view.

11.2 Verbal Reasoning

The critical reading section is based on written discourse. Male and female references are balanced, and representative minority-relevant content is included in each test. Approximately 72% (48) of the items are based on passages, while 28% (19) of the items are in the sentence completion format.

Sentence completion items are useful for measuring an understanding of the relationships among words and concepts, an understanding of the structure of the text, and knowledge of vocabulary. Within a given form of the critical reading section, a balance exists between those items that primarily measure vocabulary and those that measure reasoning about the logic of a sentence.

The passage-based reading content is balanced across four categories: humanities, social studies, natural sciences, and literary fiction. The preponderance of the items (approximately 80%) measure higher-level reading skills of the following types:

- **Primary purpose:** These questions ask about the main idea of a passage or about the author's primary purpose in writing the passage. They address the passage as whole, or an entire paragraph, rather than focusing on a smaller part of the passage. These questions tap both the process of understanding discourse and of interpreting discourse.
- **Rhetorical strategies:** These questions usually focus on a specific part of a passage—often on a particular word, image, phrase, example, or quotation—and ask why this particular element is present or what purpose it serves, rather than simply on what it means. Such questions involve the processes of interpreting discourse and evaluating discourse.
- **Implication and evaluation:** These questions go beyond the passage by asking what the information presented in the passage suggests, or what can be inferred about the author's view. They might also ask the test taker to evaluate ideas or assumptions in a passage, or to evaluate the relationship between a pair of passages. These questions involve the process of evaluating the discourse and may involve aspects of creating new understandings.
- **Tone and attitude:** These questions ask about the author's tone or attitude in the whole or a specific part of the passage. Such questions tap into the test taker's ability to interpret discourse.
- **Application and analogy:** These questions may address a specific idea or relationship in a passage and ask the test taker to recognize a parallel idea or relationship in a different

context. Such questions may also ask the test taker to recognize an additional example that would support an idea presented in the passage or may ask about an analogy that is used. Alternatively, these questions may ask how ideas presented in one passage apply to another passage, or how the author of one passage would be likely to react to an idea expressed in a related passage. Such questions draw on the test taker's ability to evaluate discourse and to create new understandings.

A few questions in each critical reading section test the literal comprehension of what is being said in a particular part of the passage. A few others—known as vocabulary in context questions—probe what a specific word means as it is used in a passage. Both of these question types draw on the process of understanding discourse.

The critical reading section taps several of the underlying dimensions posited by Burton, Welsh, Kostin, and Van Essen (2004), especially the breadth and depth of understanding in a receptive mode. The critical reading section samples the construct of verbal reasoning in a variety of ways. The detailed specifications (see Tables 2-1 through 2-3) ensure that each succeeding form or version of the test samples similar aspects of that construct. In addition, key aspects of the process of communicating are addressed in the writing portion of the SAT (see Section 11.4).

11.3 QUANTITATIVE REASONING

Dwyer, Gallagher, Levin, and Morley (2003) have reviewed the research on quantitative reasoning in an effort to better define the construct for assessment purposes. They observe, "Although the assessment of quantitative reasoning has been a measurement goal from early in the 20th century, systematic treatment of quantitative reasoning as a cognitive process distinct from mathematics as content or curriculum did not begin to take shape until much later" (p. 7). Further, they point out "that it is critical to the interpretation of reasoning tests to differentiate between elements of the reasoning construct itself that is the target of the assessment and the common core of content knowledge that all test takers are assumed to bring to the test" (p. 12). They recognize that "it is not possible, however, to assess quantitative reasoning without the content since it is the manipulation and application of the content that allows test takers to demonstrate their reasoning" (p.13). Dwyer et al. define quantitative reasoning "as the ability to analyze quantitative information" and note that it includes six capabilities:

1. Reading and understanding information given in various formats, such as in graphs, tables, geometric figures, mathematical formulas or in text
2. Interpreting quantitative information and drawing appropriate inferences from it
3. Solving problems using arithmetical, algebraic, geometric, or statistical methods
4. Estimating answers and checking answers for reasonableness

5. Communicating quantitative information verbally, numerically, algebraically, or graphically

6. Recognizing the limitations of mathematical or statistical methods (p.13)

Dwyer et al. (2003) stress that the validity and fairness of an assessment of quantitative reasoning depends on limiting the content of the assessment to a level of mathematical knowledge that is explicitly assumed to be common throughout the testing population (p.15). Independent of any particular mathematical content or level of mathematical achievement, Dwyer et al. posit a problem-solving process of three multifaceted steps:

1. Understanding and defining the problem

2. Solving the problem

3. Understanding results

This problem-solving process becomes the target for any assessment of quantitative reasoning even though the authors acknowledge, “in practice, most tests are designed to assess only a portion of the quantitative reasoning process” (Dwyer et al., 2003, p.15). In responding to the SAT mathematics questions, students need to apply this process in the context of two different item types—multiple-choice questions and student-produced responses—in which a student must solve the problem and fill in the numeric response (no options are provided). There are 44 items in multiple-choice format and 10 in student-produced-response format.

Students must apply this problem-solving process to questions drawn from a particular content domain within mathematics. In broad terms, they must have knowledge of numbers and operations, algebra and functions, geometry, measurement, statistics, probability, and data analysis. The test appropriately includes such content from a third-year high school mathematics course exponential growth, absolute value, and functional notation. The test also places emphasis on other topics, such as linear functions, manipulations with exponents, and properties of tangent lines.

Two aspects of the SAT underscore that this is a test of quantitative reasoning rather than solely mathematical knowledge: (1) students are permitted to use a four function, scientific, or graphing calculator on the test—although it is possible to solve every question without a calculator; and (2) students are provided with commonly used formulas in the test book itself, so that they do not have to memorize them. The purpose of these two “helps” is to send a clear signal to the test taker about the reasoning nature of the test.

The specifications for the mathematics section of the SAT were presented in Chapter 2, Tables 2-9 through 2-12. Each form of the test is defined in terms of the item types to be used, the mathematical content that provides the opportunity for demonstrating quantitative reasoning, as well as the distribution of questions of different levels of difficulty.

11.4 WRITING

The SAT writing test includes a direct measure of writing proficiency. Writing is an extremely complex activity: it can include different modes of discourse (e.g., narration, argumentation, description), while calling on a range of cognitive skills (e.g., interpreting, analyzing, synthesizing, organizing) and requiring various kinds of knowledge (e.g., understanding linguistic structures). Thus, it is not useful to think of writing as a unitary construct. Breland, Bridgeman, and Fowles (1999) observe, “Even if a unitary construct of writing could be defined, no single test could possibly assess the full domain” (p. 1).

On the SAT writing test, the student is asked to write a first draft essay and respond to multiple-choice questions that assess the ability to identify errors in sentences and to improve sentences and paragraphs. These skills relate closely to the cognitive operation of communication described by Burton et al. (2004). The specifications for the writing test may be found in Tables 2-4 through 2-8.

11.5 MULTIPLE-CHOICE QUESTIONS

The multiple-choice questions assess how well students use standard written English and test students’ ability to identify sentence errors, improve sentences, and improve paragraphs. The multiple-choice writing questions are used to evaluate a student’s ability to

- use language that is consistent in tenses and pronouns;
- understand parallelism, noun agreement, and subject-verb agreement;
- understand how to express ideas logically;
- avoid ambiguous and vague pronouns, wordiness, improper modification, and sentence fragments; and
- understand proper coordination and subordination, logical comparisons, diction, idiom, modification, and word order.

The multiple-choice writing questions do not ask the students to define or use grammatical terms and do not test spelling and capitalization. Using the multiple-choice format, the test assesses a student’s control of different levels of writing. Focused on improving sentences, some (25) questions ask the student to recognize and correct faults in usage and sentence structure, as well as recognize effective sentences that follow the conventions of standard written English. Others (18) ask the student to recognize and correct errors of grammar and usage in sentences. The third type of multiple-choice question asks the student to improve paragraphs. This type of question assesses a student’s ability to edit and revise sentences in the context of a paragraph or entire essay, organize, and develop paragraphs in a coherent and logical manner, while applying the conventions of standard written English (College Board, 2011, pp. 22–25).

11.6 Essay Question

The SAT writing test provides 25 minutes for a student to write a first draft essay in response to an assignment question. The student is presented with a short paragraph adapted from a published text that offers a perspective on an issue and with a question that asks for his or her point of view. The student is asked to think critically about the issue and develop a point of view, using reasoning and examples taken from reading, studies, experience, or observation to support that point of view. The essay measures a student's ability, under timed conditions, to do the kind of writing required in most college courses—writing that emphasizes precise use of language, logical presentation of ideas, development of a point of view, and clarity of expression. SAT essay prompts are developed according to the following guidelines:

- They should be accessible to the general test-taking population, including students for whom English is not a first or best language.
- They should be relevant to a wide range of fields and interests, and neither require specialized knowledge nor give an advantage to students who have completed a specific course of study.
- They should engage high school–age students while stimulating critical reflection about important topics.
- They should be free of figurative or technical language or specific literary references.
- They should give the students the opportunity to use a broad spectrum of experiences, learning, and ideas to support their points of view.

The elements of writing that can be assessed through this direct measure are reflected in the scoring guide that Readers use to evaluate and score the student essays holistically. The scoring guide used by the Readers is displayed in Chapter 2.

11.7 How Do SAT Scores Relate to College Performance?

Much of the empirical evidence for the validity of the SAT is based on analyses of the relationship of test scores to performance in college (Angoff, 1971; Wilson, 1983; Donlan, 1984; Willingham, Lewis, Morgan, and Ramist, 1990; Hezlett, Kuncel, Vey, Ahart, Ones, Campbell, and Camara, 2001; Young and Kobrin, 2001). Drawing heavily on the Young and Kobrin review, evidence gathered since 1994 is presented below.

Kobrin and Michel (2006) explored the question of whether the SAT or high school grade point average (HSGPA) is a better predictor of college freshman grade point average (FGPA) for students with high FGPAs compared to students with lower FGPAs. Employing logistic regression, they predicted the probability of a student successfully achieving a FGPA at various levels, based on that student's SAT scores and HSGPA. They found that in the total sample, at all success criterion levels except the 2.5 level, the SAT was equal to or slightly more accurate than HSGPA in predicting successful students, but generally less accurate than HSGPA in predicting unsuccessful students. However, at the highest FGPA level, 3.75 or higher, neither the SAT nor the HSGPA was able to predict successful students. Across each of the racial/ethnic groups, the SAT was typically a better predictor of successful students, and HSGPA was typically a better predictor of unsuccessful students. For students attending the most selective colleges, the SAT was more effective than or equally effective as HSGPA in predicting success at nearly all FGPA criterion levels. However, for students attending the least selective colleges, HSGPA tended to be a better predictor of success.

Norris, Oppler, Kuang, Day, and Adams (2006) studied the predictive and incremental validity of a prototype version of the recently introduced SAT writing section. Data were collected in 2003–2004 from 13 institutions, both public and private, from different sections of the country. The study included institutions of different levels of selectivity and of different size freshman classes. Data were available for a total of 1,572 students who took the SAT writing prototype and who also took the operational SAT. Note that the SAT verbal (SAT-CR) and SAT mathematics (SAT-M) scores were earned in a standard administration with high motivation, whereas the writing score was earned in an experimental administration with only an unspecified monetary incentive. The incremental validity could be different if all three scores had been earned under the same motivational condition. Such data should become available in the near future.

Norris et al. (2006) obtained two criteria—FGPA and English composition grade point average (ECGPA). Because of the variability across participating institutions, all analyses were conducted within each institution, and then weighted averages were calculated and pooled across institutions to derive the overall estimate. Statistical procedures to correct for multivariate range restriction (Lord and Novick, 1968) and shrinkage (Rozeboom, 1978) were applied.

The relationship of each of the predictors with FGPA and ECGPA is shown in Table 11-1. The values in the table represent the weighted-average validity coefficients across all of the participating institutions.

Table 11-1. 2013–14 MHSA: SAT Component—Weighted Average Correlations for Predictors with FGPA and ECGPA

Predictor	*FGPA*			*ECGPA*		
	N	*Corrected*	*Uncorrected*	*N*	*Corrected*	*Uncorrected*
SAT critical reading	1,248	0.49	0.32	891	0.30	0.20
SAT mathematics	1,248	0.47	0.29	891	0.23	0.10
SAT total	1,248	0.51	0.35	891	0.28	0.17
SAT essay	1,248	0.20	0.16	891	0.18	0.14
SAT multiple-choice	1,248	0.45	0.30	891	0.31	0.22
SAT writing total	1,248	0.46	0.32	891	0.32	0.24
HSGPA	1,248	0.43	0.38	891	0.35	0.32

Note: Corrected for multivariate range restriction (Lord and Novick, 1968). Source: Norris et al. (2006), Table 9.

These data show very similar corrected correlations with FGPA for each of the section scores and HSGPA. In other words, SAT writing (SAT-W) is about as strongly related to freshman performance as are SAT-CR, SAT-M, and HSGPA. The SAT-W total, the writing multiple-choice section, as well as the SAT-

CR are fairly predictive of English composition grades with corrected validity coefficients of 0.32, 0.31, and 0.30, respectively.

11.8 Performance over Multiple Time Periods

Working as a research consortium along with four-year colleges and universities, the College Board created a national higher education database (College Board, 2006, 2007, 2008) with the primary goal of validating the SAT, which was revised in March 2005 and consists of critical reading (SAT-CR), mathematics (SAT-M), and writing (SAT-W) for use in college admissions. The first sample examined was the first-time, first-year students entering college in fall 2006, with 110 institutions providing students' first-year coursework, grades, and retention to the second-year. Mattern et. al. (2008) examined the differential validity and prediction of the SAT using a nationally representative sample of first-year college students admitted with the revised version of the test. Their findings demonstrated that there are similar patterns of differential validity and prediction by gender, race/ethnicity, and best language subgroups on the revised SAT compared with previous research on older versions of the test (see Young , 2001, for a review).

Kobrin, et. al. (2008) presents the results of a large-scale national validity study on the SAT and documents the methods undertaken to recruit institutions, collect and prepare data for analysis, and the statistical methods applied to the data. Results show that the changes made to the SAT in March 2005 did not substantially change how well the test predicts first-year college performance. The recently added writing section was found to be the most predictive of the three individual SAT sections. The best combination of predictors of first-year college grade point average (FYGPA) is high school grade point average (HSGPA) and SAT scores.

Mattern and Patterson (2009) examined the relationship between scores on the SAT and retention to the second-year of college using student level data from the freshman class of 2006 at 106 four-year institutions. Results indicate the SAT predicts second-year retention, with 95.5 percent of high performers returning but only 63.8 of low performers. Patterson, Mattern, and Kobrin (2009) replicated the Mattern et. al. (2008) and Kobrin et. al (2008) studies using data from 159,286 first-time, first-year students that enrolled in the fall of 2007. The results of the 2009 study were largely the same as the original studies conducted in 2008 on the 2006 cohort of students. All three of the previously mentioned studies were replicated by Patterson and Mattern (2011) using the 2008 cohort of students. The 2008 cohort included in the study contained 173,963 first-time, first-year students that enrolled in the fall of 2008. Results were again largely consistent with the earlier studies. SAT scores were found to be correlated with FYGPA (r=0.54), with a magnitude similar to HSGPA (r=0.56). The best set of predictors of FYGPA remains SAT scores and HSGPA (r=0.63), as the addition of the SAT sections to the correlation of HSGPA alone with FYGPA leads to substantial improvement in the prediction.

Mattern and Patterson (2010a, 2011b) followed the 2006 cohort of students into the second and third years of college to study the validity of the SAT for predicting second-year and third-year grades, respectively. The studies investigated the predictive validity of the SAT for predicting cumulative GAP and grade point average in the second- and third-year of college. Results indicate the SAT is strongly correlated with both second- and third-year outcomes.

Mattern and Patterson (2011a) replicated their earlier study (Mattern and Patterson, 2009) using data from the 2007 cohort. The results were largely the same showing that SAT scores are related to second-year retention. After controlling for student and institutional characteristics, returners had higher SAT total scores than non-returners, by an average of 116 points. This held true even within each subgroup analyzed, meaning the SAT performance gap is not due to differences in the demographic characteristics of the two groups. Additionally, differences in retention rates by student subgroups are minimized and in some instances eliminated when controlling for SAT performance. This is particularly noticeable with respect to differences in retention rates by ethnicity.

Mattern and Patterson (2010b) follows the 2006 cohort of students into their third year and replicates their 2009 study to investigate retention rates. Results indicate that SAT performance is related to third year retention rates and mirror the findings of Mattern and Patterson (2011a).

11.9 Differential Validity for Subgroups

A considerable amount of research in the last fifteen years has examined the question of whether SAT scores, as well as other predictors, have differential validity for various subgroups of the test-taking population. In other words, is there a different relationship between the predictors and the criterion of college grades for men than for women, or among members of different racial or ethnic groups? Ramist, Lewis, and McCamley-Jenkins (1994) analyzed a database of course grades from 38 colleges and universities to determine if group differences occurred in the prediction of individual course grades as well as FGPA. This was the same database that was used in the earlier study by Ramist, Lewis, and McCamley (1990). A sample of over 46,000 students was used to investigate differences by gender and by five ethnic/racial groups (Native American, African American, Hispanic, Asian American, and White). The uncorrected and corrected correlations with FGPA and with a course grade criterion (adjusted for the grading difficulty of the courses) are shown in Table 11-2. Since the total sample for Native American students was only 184, results for this group should be considered tenuous at best.

The courses taken by these students in their first year of college were assigned to 37 categories based on subject, skills required, and level. For example, there were five categories for mathematics (based on level) and nine for English (based on level as well as whether the emphasis was on reading/literature, writing/composition, or both). Their results showed differences in course-taking behavior for the different gender and ethnic/racial groups.

Table 11-2. 2013–14 MHSA: SAT Component—Effectiveness by Student Group Correlation with FGPA

		Gender		*Ethnic Group*				
N	*All Students*	*Male*	*Female*	*Native American*	*Asian American*	*African American*	*Hispanic*	*White*
	46,379	22,412	23,967	184	3,848	2,475	1,599	36,743
	Correlations With FGPA*							
SAT-CR	0.50	0.48	0.55	0.42	0.47	0.44	0.39	0.50
SAT-M	0.53	0.53	0.58	0.36	0.56	0.44	0.38	0.52
SAT (V+M)	0.57	0.56	0.62	0.49	0.58	0.49	0.43	0.56
HSGPA	0.61	0.58	0.61	0.49	0.60	0.46	0.53	0.61
V+M+H	0.68	0.65	0.71	0.63	0.69	0.56	0.58	0.68
	Correlations* With Course Grade Criterion							
SAT-CR	0.50	0.48	0.53	0.39	0.49	0.47	0.44	0.49
SAT-M	0.54	0.53	0.57	0.32	0.59	0.48	0.48	0.53
SAT (V+M)	0.60	0.59	0.64	0.48	0.63	0.57	0.55	0.59
HSGPA	0.58	0.57	0.59	0.59	0.63	0.46	0.55	0.57
V+M+H	0.70	0.69	0.74	0.70	0.76	0.64	0.68	0.69

* Correlations corrected for restriction of range and criterion unreliability. Source: Ramist, Lewis, and McCamley-Jenkins (1994), Tables 1 and 4

11.9.1 Gender

Drawn from the Ramist et al. (1994) study, Table 11-2 shows that the correlations between the predictor variables and both the FGPA and the course grade criteria were higher for females than for males, more so for the SAT than for HSGPA, and more so for the verbal score than for the mathematics score. For both criteria, the correlation of HSGPA exceeded the correlation of the combined SAT-CR and SAT-M for males, but for females, the SAT showed a stronger correlation than did HSGPA. Using both HSGPA and SAT scores, the corrected correlation for predicting FGPA was higher for females (0.71) than for males (0.65), as was the corrected correlation for predicting course grade (0.74 versus 0.69).

In a 1994 report, Pennock-Román investigated gender differences in the prediction of college grades at four universities: two in California, one in Massachusetts, and one in Texas. As in the Ramist et al. (1994) study, Pennock-Román found that males were more likely to take courses in the physical sciences and engineering, while females were more likely to take courses in the humanities and social sciences.

Since it has been widely observed at many institutions that the average grade earned by students in courses varies considerably from department to department, one explanation for the underprediction of women's grades is that this is due to differences in course selection. Because it is more common for women to enroll in courses where the average grade is higher than in the courses that men take, the underprediction of women's grades may result from differences between men and women in the courses used to compute FGPA or CGPA. Pennock-Román (1994) sought to examine this hypothesis by developing and using a variable (MAJSCAL) that reflected the "degree of grading toughness" of the student's category of college major. Separate prediction equations, by sex, of FGPA from SAT scores and HSGPA were used to calculate MAJSCAL. The average residual for the students who majored in a given department was used as an indication of the “grading toughness” of that department. The magnitude of the residual for each department was then converted to the ordinal scale used for MAJSCAL.

The FGPAs of women were underpredicted using all predictors (HSGPA, SAT verbal, SAT mathematics, or all three combined) at all four universities. This finding was also true for three subgroups of women (Asian American, White, and a combined group of African American and Latino students), with the exception of Asian American female students at the Texas university. For example, when all three predictors were used, the average underprediction of women’s grades ranged from 0.019 for Asian American females at one of the California schools to 0.185 for White females at the Texas institution. When MAJSCAL was used as an additional predictor, the underprediction of women’s FGPAs was significantly reduced but not completely eliminated. This study provided further evidence that gender differences in the selection of college courses and majors may be the main reason behind the underprediction of women’s grades. The use of MAJSCAL, a measure that is relatively easy to construct and understand, substantially reduced the degree of underprediction. In addition, by incorporating information on college majors through a measure such as MAJSCAL, a reasonable, practical procedure for controlling departmental grading differences may be available for use in future studies of differential prediction.

The recent study by Bridgeman, McCamley-Jenkins, and Ervin (2000) examined the impact of changes to the content and scale of the SAT on the predictive validity of the SAT overall as well as for subgroups of students. Results indicated that the correlations of SAT verbal, SAT mathematics, and SAT composite with FGPA, averaged across all the schools, were higher by 0.03 to 0.05 for women than for men. The average correlation of HSGPA with FGPA was slightly higher (by 0.02 to 0.03) for men than for women. When less-selective institutions were analyzed separately, these correlations were found to be higher for females. Other studies of differential validity that have examined data from highly selective institutions have also found that gender differences in validity are often smaller than at less-selective institutions (Ramist et al., 1994).

The combination of SAT score and HSGPA was about equally effective in predicting FGPA for men (multiple correlation of 0.44) and for women (0.45). At the most selective institutions (with an average SAT composite score over 1250), the grades of men and women were predicted equally well. In contrast, at schools with lower average SAT scores, the grades of females were more predictable than the grades of males. As with other studies of differential prediction, Bridgeman et al. (2000) found that the grades of women were underpredicted from SAT scores alone (with an average underprediction of 0.11); from SAT scores and HSGPA (0.07); and from SAT scores, HSGPA, and an adjustment factor for course difficulty (0.05).

In Young and Kobrin’s review (2001) of the literature on differential validity and prediction with regard to gender differences, the correlations between predictors and criterion were generally higher for women than for men. In terms of prediction, the typical finding in these studies was that women’s college grades were underpredicted. However, in the most selective universities, the correlations for men and women appeared to be equal, and the degree of underprediction for women’s grades appeared to be noticeably less than at other institutions. Compared with earlier studies on this topic, gender differences in validity and prediction appear to have persisted, although the magnitude of the differences seems to have recently decreased.

11.9.2 Race and Ethnicity

In the Ramist, Lewis, and McCamley-Jenkins (1994) study reported in Table 11-9, the highest correlation of SAT-CR with FGPA was for White students (0.50) and the lowest was for Hispanic students (0.39). For SAT-M, the lowest correlation was for Native American students (0.36) and the highest was for Asian American students (0.56). This may reflect the fact that Asian American students took more quantitatively oriented courses than the other subgroups, a fact confirmed by Bridgeman, Pollack, and Burton's (in press) *Predicting Grades in Different Types of College Courses*. Asian American students had the highest multiple correlation for test scores combined with HSGPA (0.69), while African American students had the lowest (0.56). Results for predicting individual course grades were comparable to those for predicting FGPA, with the highest corrected correlations for the combination of SAT-CR, SAT-M, and HSPGA for Asian American (0.76), Native American (0.70), and White (0.69) students. For four of the five ethnic groups, the combination of SAT-CR and SAT-M scores was equal to or better than HSGPA in predicting course grades.

Both FGPA and course grades of Native American, African American, and Hispanic students were overpredicted; that is, they earned lower grades in college than was predicted, using any predictor, alone or in combination, while the grades of Asian American students were underpredicted. The magnitude of the overprediction was largest for Native American, followed by African American, and finally Hispanic students. Performance for Native American students was overpredicted in a variety of science, foreign language, English, and mathematics courses; African American student performance was overpredicted, especially in quantitative and science courses; Hispanic student performance was overpredicted in most courses. Course performance of Asian American students was underpredicted in mathematics and science but overpredicted in English, architecture, and physical education. The performance of White students was slightly underpredicted in English and overpredicted in mathematics and technical/vocational courses.

The Bridgeman, McCamley-Jenkins, and Ervin (2000) study found that correlations of SAT-CR, SAT-M, and SAT composite with FGPA were uniformly higher for women than for men in the four subgroups studied (African American, Asian American, Hispanic, and White). However, the results for HSGPA were mixed, with some correlations higher for one gender or the other, depending on the ethnic/racial subgroup. The combination of SAT score and HSGPA appeared to be equally effective across all of the ethnic/racial subgroups and for men and women within each subgroup. The single exception to this finding was the somewhat lower multiple correlation for Hispanic men (0.38) as compared to Hispanic women (0.44).

The differential prediction findings indicated that, using SAT score and HSGPA, the grades of women from three of the subgroups were underpredicted. On average, the largest underprediction was for

White (0.09), then Asian American (0.07), and finally African American (0.05) women. The grades of Hispanic women were slightly overpredicted at 0.02. Adding the adjustment factor served to reduce the underprediction (or increase the overprediction) by 0.01 to 0.03. For men, the largest overprediction occurred in African American (0.16), followed by Hispanic (0.12), then White (0.09) students. The grades of Asian American men were accurately predicted. Adding the adjustment factor changed the overprediction only slightly for African American, Hispanic, and White men (by 0.02 or less), but caused the grades of Asian American men to be underpredicted by 0.05.

In 2001, Young and Kobrin produced a comprehensive review and analysis of all of the available differential validity and prediction studies published between 1974 and 2000. (See also Young [2004] for a further discussion of these differential validity and prediction studies.) In all, 29 studies of ethnic/racial differences and 37 studies of gender differences were reviewed. Young and Kobrin provided detailed information on each of the studies in the review, including type of study, name of institution(s), specific cohorts, sample sizes, predictors and criterion used, and values of validity coefficients and prediction results reported by each study's author(s). In addition, a short descriptive summary of each study was included. In another section of the report, Young summarized the findings from five earlier research reviews on differential validity and prediction (Breland, 1979; Duran, 1983; Linn, 1973; Linn, 1982; Wilson, 1983).

With regard to ethnic and racial differences, Young and Kobrin (2001) reported that the subgroups that have been studied include Asian American, African American, Hispanic, and Native American students. Some studies used a combined sample of minority students composed primarily of African American and Hispanic students. Overall, there was no common pattern to the results for validity and prediction for the different subgroups. Correlations between predictors and criterion were different for each subgroup, with generally lower values for African American and Hispanic students and similar values for Asian American students compared to White students. Too few studies of Native American or of combined samples of minority students were available to reliably determine typical validity coefficients for these groups. In terms of grade prediction, the common finding was one of overprediction of college grades for all minority groups with the exception of Asian American students, although the magnitude differed for each group. With Asian American students, studies that adjusted grades to account for differences in course difficulty found that grades were underpredicted.

11.9.3 Students with Disabilities

Increased attention to testing procedures for students with disabilities occurred in 1977 when the U.S. Department of Education issued regulations implementing Section 504 of the Rehabilitation Act of 1973. The regulations require individualized testing accommodations, validation of admissions tests for examinees with disabilities, and assurance that the tests are measuring aptitude and achievement without the impact of extraneous variables attributed to disability (Willingham, Ragosta, Bennett, Braun, Rock, and Powers, 1988). In response to Section 504, the College Board and the ETS sponsored a four-year study that focused primarily on students with different disabilities who had taken admissions testing program exams. Data on score reliability and validity did not show dependable differences in precision between students with disabilities and those without (Bennett, Ragosta, and Strickler, 1984). For most students with disabilities, the combination of high school grades and test scores remained the best predictor of college performance. Some exceptions noted were an underprediction of college freshman grades for deaf or hearing impaired students, an overprediction for students with specific learning disabilities, and a slight overprediction for students with physical disabilities.

Other studies investigated the validity of SAT scores for examinees with and without disabilities. Bennett, Rock, and Kaplan (1985) examined verbal and mathematics scores for groups of examinees (with and without disabilities) to discern whether SAT scores were comparable across individuals tested under standard administration procedures versus those tested under special administrations (including extended time). Findings suggested that SAT scores are generally equally reliable and valid for predicting the performance of students with and without disabilities. Similarly, Ragosta, Braun, and Kaplan (1991) tested the validity of SAT scores for predicting overall performance and persistence of college students with disabilities and found that scores were a good predictor of both variables.

Extended Time Accommodations

Students with specific learning disabilities comprise approximately 90% of examinees who request accommodations on the SAT (Camara and Schneider, 2000) and account for the largest percentage of college freshmen with disabilities (Cahalan, Mandinach, and Camara, 2002). In addition, extended time is the most often requested and granted accommodation on college admissions tests. As such, more recent studies have focused on students with specific learning disabilities who take the SAT with extended time to determine the impact that providing extra time may have on performance.

Providing extended time accommodations for SAT examinees with documented disabilities is based on the notion that test timing is a primary source of noncomparability between test scores (i.e., certain disabilities may lead to slower processing of test content). Data from test administration timing records were used to establish empirically derived testing times for special administrations of the SAT for examinees with disabilities and to establish eligibility guidelines for individuals requesting special administrations (Ragosta and Wendler, 1992). This research established that comparable testing time for students with disabilities was between 1.5 and 2 times that for students without disabilities. These time limits assured that approximately equal percentages of students from both groups would complete each section of the SAT. An exception was students with visual impairments or blindness using Braille or cassette versions of the test, who required between double and triple the normal time limits.

Camara, Copeland, and Rothschild (1998) examined the impact of extended time on SAT performance. They compared the mathematics and verbal section score gains for students who received an extended time accommodation and completed each SAT section in standard time (75 minutes), up to time and a half (an additional 1 to 38 minutes), time and a half to double time (an additional 39 to 75 minutes), and greater than double time (an additional 76 or more minutes). Findings revealed that time and a half to double time produced the highest score gains on the mathematics section, and greater than double time produced the highest score gains on the critical reading section.

In a study on the effects of taking the SAT with extended time for students with specific learning disabilities, Camara and Schneider (2000) cited important conclusions about extended time administrations. One conclusion is that allowing students to retest using extended time does lead to SAT score improvement, but the amount of improvement is modest. Average score gains with extended time are 32 points on the verbal scale and 26 points on the mathematics scale. Overall, there is a positive correlation between the amount of extended time allowed and the amount of score gain. While extended time does enable students with learning disabilities to perform better on the SAT, the standard allowance of time and a half or double time may overcompensate for some students and result in overprediction of college performance. Finally, the study found that students who scored higher on their initial SAT examination used more time in a subsequent administration and experienced larger score gains than their peers who received lower scores on the initial examination.

Cahalan, Mandinach, and Camara (2002) examined the predictive validity of scores from the SAT for students who received special testing accommodations. Particularly, they were interested in students with specific learning disabilities who had taken the SAT between 1995 and 1998 with an extended time accommodation. The study provided evidence that scores from the SAT are a valid tool for helping admissions officers select students with specific learning disabilities (who received extended time accommodations) for college admission. While SAT scores alone are a good predictor of FGPA, the prediction is increased by using both SAT scores and HSGPA.

Morgan and Huff (2002) compared the reliability and dimensionality of the SAT critical reading and mathematics sections for examinees tested under standard timing conditions and examinees tested with extended time accommodations. Four comparisons were conducted between the standard time and extended time groups for May 2001 critical reading and mathematics and October 2001 critical reading and mathematics. Reliability and standard error of measurement estimates across the two groups of examinees differed slightly for all four comparisons, with the extended time group showing slightly more measurement error than the standard time group. Results from item-level factor analyses and multidimensional scaling analyses produced no evidence to suggest that the scores on the SAT I have different interpretations when the examinees have an extended time administration compared to the standard.

Lindstrom (2006) used data from the initial administration of the new SAT (administered March 17, 2005) to analyze a sample of 4,952 examinees. First, confirmatory factor analysis was used to assess the fit of a single-factor structure model for the mathematics, critical reading, and writing sections to each of the two groups. Next, a study of factorial invariance examined whether a common factor model for the mathematics, critical reading, and writing sections holds across the two groups at increasingly restrictive levels of constraint. Invariance across the two groups was supported for factor loadings, thresholds, and factor variances. Thus, there was no real evidence to suggest that the scores on the mathematics, critical reading, and writing sections of the SAT have different interpretations when examinees have an extended time administration as opposed to the standard time administration.

11.9.4 Fatigue Effects

Cahalan-Laitusis, Morgan, Bridgeman, Zanna, and Stone (2007) examined operational data from the SAT to determine if students who tested under extended time conditions were suffering from excessive fatigue relative to students who tested under standard time conditions. Excessive fatigue was defined by significant increases in differential item functioning (DIF) and decreases in item completion rates, for items at the end of testing compared to the beginning of testing. Both of these factors were examined by comparing the performance of students who tested under standard time to students testing with extended time on items administered early in the test (Sections 2 or 3) and different items administered late (Sections 8, 9, or 10) during the 10-section test administration. Results indicated few changes in the level of DIF. In addition, item completion rates for students who received extra time were comparable to test takers without disabilities who tested under standard time on both early and late sections.

11.10 Summary of the MHSA SAT Component

This section began with a discussion of what is measured by the SAT. The substance of the test represents a complex interaction between the particular reading, mathematical, and writing skills; the content through which students are asked to demonstrate their skills; and the types of questions used to elicit that demonstration of skills. The test does not include esoterica, but rather focuses on the application of content and skills that are part of a typical high school experience.

The second portion of the section reviewed evidence of the relationship of the substance of the test to what teachers judge to be important in each domain, and the intensity with which it is treated in the classroom. The third portion of the section reported on evidence demonstrating that the scores on the revised (2005) SAT can be interpreted in the same way as earlier scores and argued that the predictive validity evidence collected over past decades can be used to support the interpretation of the revised test.

The final portion of the section examined the relationship of SAT scores to performance in college, as measured by different criteria such as freshman GPA, four-year cumulative GPA, college graduation, or performance in an English composition course. Research on the differential validity of the test by gender and racial/ethnic group was also presented.

Overall, there is a substantial body of evidence that supports the use of the SAT in the college admissions process. Even within homogeneous groups with similar high school preparation and grades attending a particular stratum of colleges, the SAT differentiates between those who are academically more successful and those who are less so. The SAT does not account for all the variation in college performance, but it does provide a good indicator of how a student is likely to perform in the particular context of a college or university.

CHAPTER 12 THE MHSA SCIENCE COMPONENT

Because interpretations of test scores, and not a test itself, are evaluated for validity, the purpose of the *2013–14 MHSA Technical Report* is to describe several technical aspects of the MHSA in support of score interpretations (AERA et al., 1999). Each chapter contributes an important component in the investigation of score validation: test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

The MHSA science test, as described in Chapters 3 and 5, was written and aligned in its entirety to Maine's Learning Results (MLRs) science accountability standards. MHSA science results are intended to facilitate inferences about student achievement on the science standards, which in turn serve the evaluation of school accountability and inform the improvement of programs and instruction.

Standards for Educational and Psychological Testing (AERA et al., 1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. The evidence around test content, response processes, internal structure, relationship to other variables, and consequences of testing speaks to different aspects of validity but are not distinct types of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

Evidence on test-content validity is meant to determine how well the assessment tasks represent the curriculum and standards for each content area and grade level. Content validation is informed by the item development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through the lens provided by the standards, evidence based on test content was extensively described in Chapter 3. Item alignment with accountability standards; item bias, sensitivity, and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training are all components of validity evidence based on test content. As discussed earlier, all questions are aligned by Maine educators to the 2007 MLRs and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (constructed-response and multiple-choice). Finally, tests are administered according to state-mandated standardized procedures, with allowable accommodations.

Chapter 5 provided additional content validation evidence in describing mandated standardized testing procedures, including the requirement that all test coordinators and test administrators familiarize themselves with and adhere to the procedures outlined in the *Principal and Test Coordinator Manual* and *Test Administrator Manual*. The quality control procedures related to scanning and machine scoring, as well as the training and monitoring of readers, presented with the scoring information in Chapter 7 added to the body of content validation evidence.

Evidence based on internal structure is presented in great detail in the discussions of item analyses, reliability, and scaling and equating in Chapter 9. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), differential item functioning (DIF) analyses, dimensionality analyses, reliability, standard errors of measurement (SEMs), and item response theory (IRT) parameters and procedures. In general, item difficulty and discrimination indices were in acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall.

Evidence based on the consequences of testing is addressed in the scaled-score information in Chapter 9 and the reporting information in Chapter 10. Each of these chapters speaks to the efforts undertaken to promote accurate and clear information provided to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across subsequent years. Achievement levels provide users with reference points for mastery, which is another useful and simple way to interpret scores. Several different standard reports are provided to stakeholders.

12.1 QUESTIONNAIRE DATA

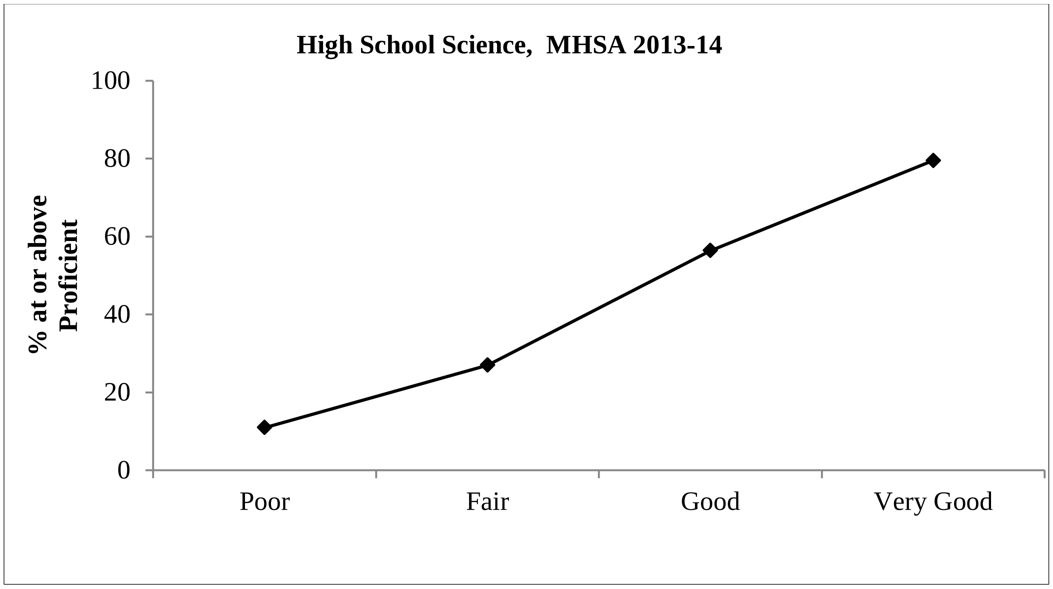
External validity of the MHSA is conveyed by the relationship of test scores and situational variables such as self-image, attitude toward content matter, and match of test questions to what is learned in school. These situational variables were all based on student questionnaire data collected during the administration of the MHSA. Note that no inferential statistics are included; however, because the numbers of students are quite large, differences among average scores may be statistically significant.

12.1.1 Self-Image

Examinees were asked how they would rate themselves as a student in science. Figure 12-1 on the following page indicates a strong positive relationship between self-image as a student and MHSA scores.

Question: Which of the following best describes how you rate yourself as a student in science?

Figure 12-1. 2013–14 MHSA Science: Questionnaire Results—Self-Image

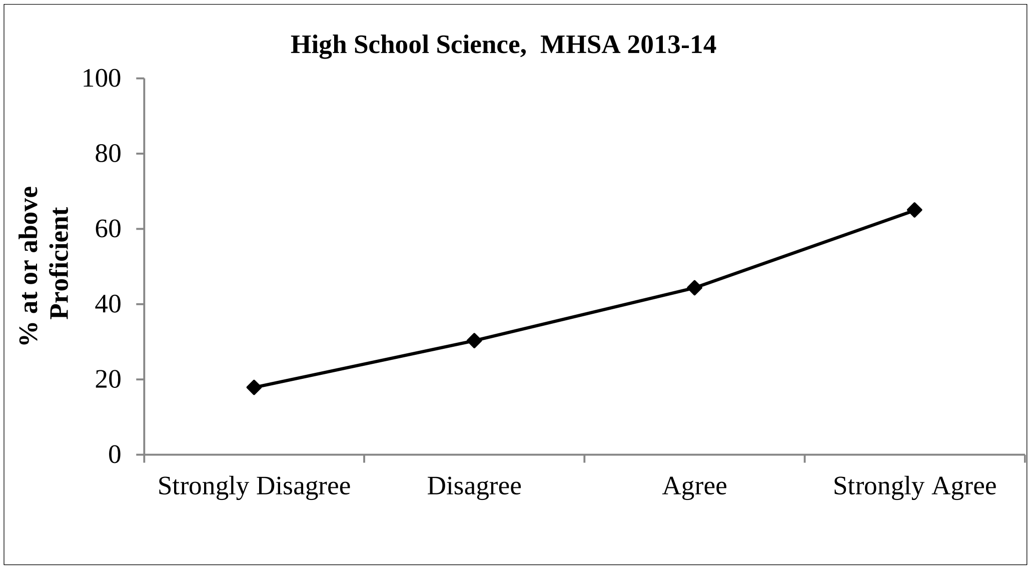


12.1.2 Attitude Toward Content Area

Students were asked how they felt about the statement “My knowledge of science will be useful to me as an adult.” Figure 12-2 indicates that students’ attitudes toward science are related positively to MHSA scores.

Statement: My knowledge of science will be useful to me as an adult.

Figure 12-2. 2013–14 MHSA Science: Questionnaire Results—Attitude

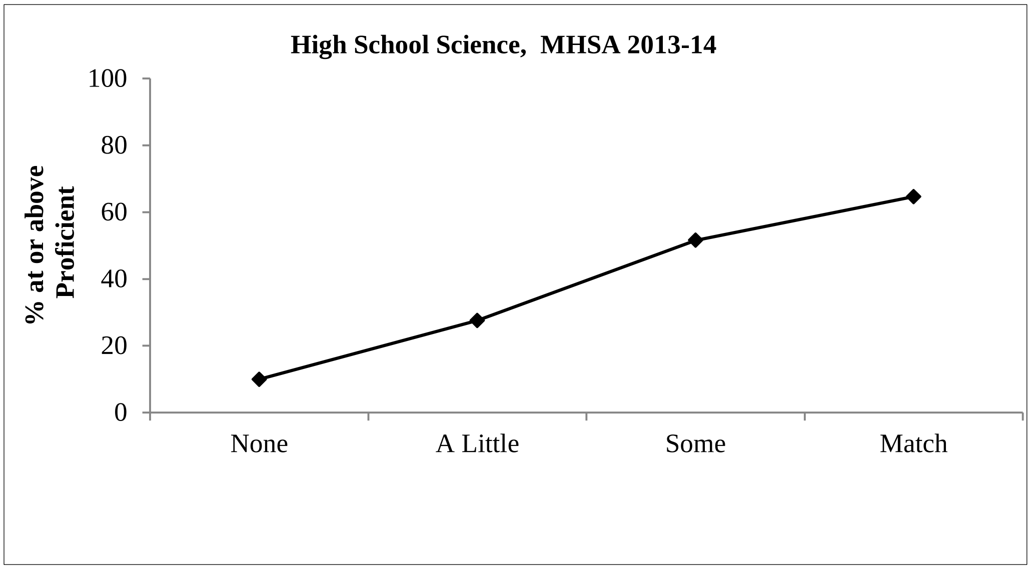


12.1.3 Match of Questions to What Is Learned in School

Students were asked how well the questions on the MHSA test matched what they had learned in school about science. Figure 12-3 indicates that there is a positive relationship between how well students feel the questions match what they have learned in science and MHSA scores.

Question: How well do the questions that you have just been given on this MHSA test match what you have learned in school about science?

Figure 12-3. 2013–14 MHSA Science: Questionnaire Results—Assessment Matches School

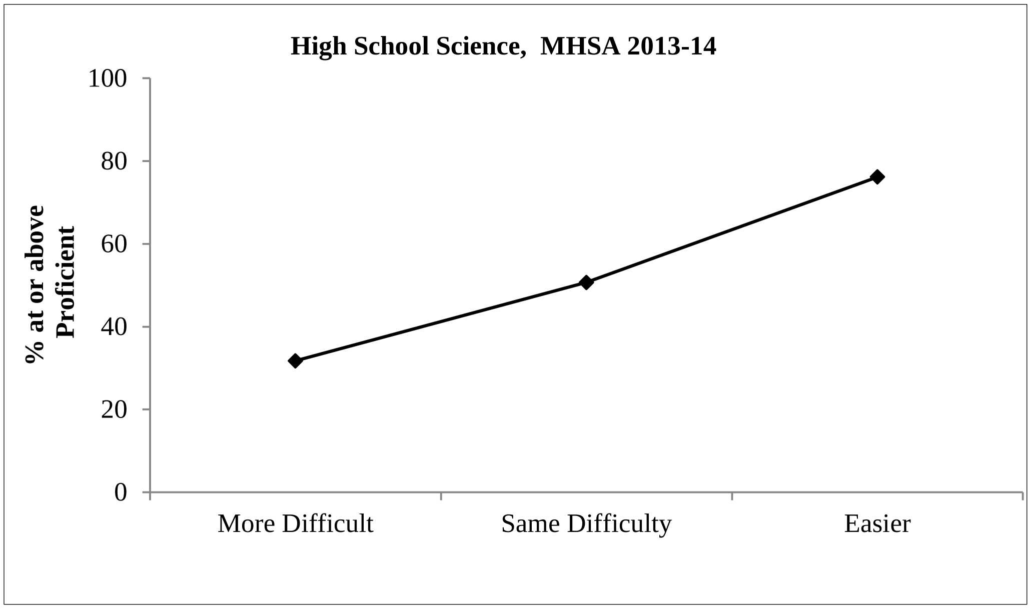


12.1.4 Difficulty of Assessment

Students were asked how difficult they found the test. Figure 12-4 on the following page indicates that there is a strong negative relationship between how difficult the students felt the items were and overall MHSA science scores (i.e., students who found the test more difficult received lower scores than students who found the test easier).

Question: How difficult was this science test?

Figure 12-4. 2013–14 MHSA: Questionnaire Results—Difficulty



The evidence presented in this report supports inferences of student achievement on the content represented in *Maine's Learning Results* and grade level expectations for science for the purposes of program and instructional improvement and as a component of school accountability.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Angoff, W. H. (Ed.). (1971). The College Board admissions testing program: A technical report on research and development activities related to the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.

Baker, F. B., & Kim, S-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bennett, R. E., Ragosta, M., & Strickler, L. (1984). *The test performance of handicapped people* (Report No. 84-32). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1985). *The psychometric characteristics of the SAT for nine handicapped groups* (ETS Research Report RR-85-49). Princeton, NJ: Educational Testing Service.

Bowen, W. G., & Bok, D. (1998). The shape of the river: Long-term consequences of considering race in college and university admissions. Princeton, NJ: Princeton University Press.

Breland, H. M. (1979). *Population validity and college entrance measures* (Research Monograph No. 8). New York: The College Board.

Breland, H. M., Bridgeman, B., & Fowles, M. (1999). *Writing assessment in admission to higher education: Review and framework* (CBR No. 99-3). New York: The College Board.

Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999). *The performance assessment study in writing: Analysis of the SAT II: Writing Subject Test* (CBR No. 99-4). New York: The College Board.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (CBRR 2000-1). New York: The College Board.

Bridgeman, B., Pollack, J., & Burton, N. (2004). Understanding what SAT Reasoning Test scores add to high school grades: A straightforward approach (CBRR 2004-4). New York: The College Board.

Bridgeman, B., Pollack, J., & Burton, N. (2008). *Predicting grades in different types of college courses.* Princeton, NJ: Educational Testing Service.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth: Holt, Rinehart and Winston.

Burton, N. W., & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (CBRR 2001-2). New York: The College Board.

Burton, N. W., Welsh, C., Kostin, I., & Van Essen, T. (2004). *Toward a definition of verbal reasoning in higher education.* Unpublished manuscript.

Cahalan, C. (2000). *Geographic clusters of learning disabled test takers in the United States.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 443 841).

Cahalan, C., Mandinach, E. B., & Camara, W. J. (2002). *Predictive validity of SAT I: Reasoning Test for examinees with learning disabilities and extended time accommodations* (CBRR No. 2002-5). New York: College Entrance Examination Board.

Cahalan-Laitusis, C., Morgan, D. L., Bridgeman, B., Zanna, J., & Stone, E. (2007). *Examination of fatigue effects from extended time accommodations on the SAT Reasoning Test* (CBRR 2007-1). New York: The College Board.

Camara, W. J., Copeland, T., & Rothschild, B. (1998). *Effects of extended time on the SAT I: Reasoning Test score growth for students with disabilities* (CBRR No. 98-7). New York: College Entrance Examination Board.

Camara, W. J., & Schneider, D. (2000). *Testing with extended time on the SAT I: Effects for students with learning disabilities* (College Board Research Note No. RN-08). New York: College Entrance Examination Board.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56,* 81–105.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21*(4), 559–566(8).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

College Board. (2004). SAT preparation booklet 2004–2005 for the new SAT. New York: Author.

College Board. (2005a). 2005 college bound seniors: Total group profile report. New York: Author.

College Board. (2005b). *The new SAT: Implemented for the class of 2006.* Retrieved January 21, 2005, from www.collegeboard.com.

College Board. (2005c). *The new SAT: A guide for admission officers*. New York: Author.

College Board. (2005d). *Report for the State of Maine on the alignment of the SAT and PSAT/NMSQT to the Maine Learning Results*. Internal report provided to the Maine Department of Education in September 2005.

College Board. (2006). 2006 College-Bound Seniors: Total Group Profile Report. New York: The College Board.

College Board. (2007). 2007 College-Bound Seniors: Total Group Profile Report. New York: The College Board.

College Board. (2008). The College Board College Handbook 2008 (45th ed.). New York: The College Board.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Crone, C. R., & Schmitt, A. P. (1991). *Alternative verbal aptitude item types: DIF issues.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Donlan, T. F. (Ed.). (1984). The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.

Dorans, N. J. (2000). *Distinctions among classes of linkages* (College Board Research Note RN-11). New York: The College Board.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1), 55–84.

Dorans, N. J. (2004a). Equating, concordance and expectation. *Applied Psychological Measurement, 28*(4), 227–246.

Dorans, N. J. (2004b). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43–68.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Dorans, N. J., Liu, J., & Hammond, S. (2004). The role of the anchor test in achieving population invariance across subpopulations and test administrations. *Applied Psychological Measurement.*

Draper, N. R. & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Duran, R. P. (1983). Hispanics' education and background: Predictors of college achievement. New York: The College Board.

Dwyer, C. A., Gallagher, A., Levin, J., & Morley, M. (2003). *What is quantitative reasoning? Defining the construct for assessment purposes* (RR-03-30). Princeton, NJ: Educational Testing Service.

French, J. W. (1957). Validation of the SAT and new item types against four-year academic criteria (RB-57-4). Princeton, NJ: Educational Testing Service.

Gulliksen, H. (1950). *Theory of mental tests*. New Jersey: Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Hezlett, S. A., Kuncel, N. R., Vey, M., Ahart, A. M., Ones, D. S., Campbell, J. P., & Camara, W. (2001). *The effectiveness of the SAT in predicting success early and late in college: A meta-analysis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Holland, P. W., & Thayer, D. T. (1988) Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (Eds.) *Test validity*, (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Joint Committee on Testing Practices (2004). *Code of fair testing practices in education*. Washington, DC: National Council on Measurement in Education.

Khaliq, S., & Reshetar, R. (2003). *Summary of testing years 1998/1999 through 2002/2003 DIF statistics for the SAT* (Research memorandum). Princeton, NJ: Educational Testing Service.

Kline, P. (1986). A handbook of test construction: Introduction to psychometric design. London: Methuen.

Kobrin, J. L., Camara, W. J., & Milewski, G. B. (2002). *The utility of the SAT I and SAT II for admissions decisions in California and the nation* (CBRR 2002-6). New York: The College Board.

Kobrin, J. L., & Michel, R. S. (2006). *The SAT as a predictor of different levels of college performance* (CBRR 2006-3). New York: The College Board.

Kobrin, J. L. & Schmidt, A. E. (2005). *The research behind the new SAT* (Research Summary RS-11). New York: The College Board.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). Validity of the SAT for Predicting First-Year College Grade Point Average (College Board Research Rep. No. 2008-5). New York: The College Board.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3,* 97–104.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lawrence, I. D., Lyu, C. F., & Feigenbaum, M. D. (1995). *DIF data on free response SAT I mathematical items.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lawrence, I., Rigol, G., Van Essen, T., & Jackson, C. (2002). *A historical perspective on the SAT 1926–2001* (CBRR 2002-7). New York: The College Board.

Lawrence, I. D., & Schmitt, A. (1994). Setting statistical specifications for the new SAT and PSAT/NMSQT. In Lawrence et al. (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM 94-10). Princeton, NJ: Educational Testing Service, 1–25.

Lindstrom, J. H. (2006). *The role of extended time on the SAT Reasoning Test for students with disabilities.* Unpublished research report completed as part of the College Board Student Grant Fellowships Program.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139–161.

Linn, R. L. (1982). Ability testing: Individual differences, prediction and differential prediction. In A.K. Wigdor & W.R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies (Part 2, pp. 335–388).* Washington, DC: National Academy Press.

Liu, J. (2004). Examination of long leg and short leg equatings for SAT verbal and math by administration for the 2002–03 testing year (Research memorandum). Princeton, NJ: Educational Testing Service.

Liu, J., Cahn, M. F., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linkage of new SAT® to old SAT across gender groups. *Journal of Educational Measurement, 43*(2), 113–129.

Liu, J., Feigenbaum, M., & Cook, L. (2004). *A simulation study to explore configuring the SAT® I: Verbal Test without analogy items* (College Board Research Report No. 2004-2, ETS Research Report RR-04-01). Princeton, NJ: Educational Testing Service.

Liu, J., Feigenbaum, M. D., & Dorans, N. J. (2003). *Equitability analysis of the new SAT to the current SAT I* (Statistical Report 2003-73). Princeton, NJ: Educational Testing Service.

Liu, J., Feigenbaum, M., & Walker, M. E. (2004). *New SAT and PSAT/NMSQT spring 2003 field trial design* (Statistical Report 2004-95). Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–198.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Mathematical Sciences Education Board. (1989). *Everybody counts: A report to the nation on the future of mathematics education.* Washington, DC: National Academy Press.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). Differential Validity and Prediction of the SAT (College Board Research Rep. No. 2008-4) New York: The College Board.

Mattern, K. D., & Patterson, B. F. (2009). Is Performance on the SAT Related to College Retention? (College

Board Research Rep. No. 2009-7). New York: The College Board.

Mattern, K. D., & Patterson, B. F. (2010a). Validity of the SAT for Predicting Second-Year Grades: 2006 SAT

Validity Sample (College Board Statistical Rep. No. 2011-1). New York: The College Board.

Mattern, K. D., & Patterson, B. F. (2010b). The Relationship between SAT Scores and Retention to the Third Year: 2006 SAT Validity Sample (College Board Statistical Rep. No. 2011-2). New York: The College Board.

Mattern, K. D., & Patterson, B. F. (2011a). The Relationship between SAT Scores and Retention to the Second Year: 2007 SAT Validity Sample (College Board Statistical Rep. No. 2011-4). New York: The College Board.

Mattern, K. D., & Patterson, B. F. (2011b). Validity of the SAT for Predicting Third-Year Grades: 2006 SAT

Validity Sample (College Board Statistical Rep. No. 2011-3). New York: The College Board.

Milewski, G., Johnsen, D., Glazer, N., & Kubota, M. (2005). A survey to evaluate the alignment of the new SAT writing and critical reading sections to curricula and instructional practices (RR 2005-1). New York: The College Board.

Morgan, R. (1994). *Effect of scale choice on predictive validi*ty. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Morgan, D. L., & Huff, K. (2002). Reliability and dimensionality of the SAT for examinees tested under standard timing conditions and examinees tested with extended time. Unpublished research conducted at the Educational Testing Service documented in a memorandum on July 15, 2002.

Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). *The College Board SAT writing validation study: An assessment of predictive and incremental validity* (CBRR 2006-2). New York: The College Board.

Oh, H., & Sathy, V. (2006). *Construct comparability and continuity in the SAT* (Statistical Report SR-2006-22). Princeton, NJ: Educational Testing Service.

Patterson, B.F., Mattern, K.D., & Kobrin, J.L. (2009). Validity of the SAT for Predicting FYGPA: 2007 SAT Validity Sample (College Board Statistical Rep. No. 2009-1). New York: The College Board.

Patterson, B. F., & Mattern, K. D. (2011). Validity of the SAT for Predicting First-Year Grades: 2008 SAT

Validity Sample (College Board Statistical Rep. No. 2011-5). New York: The College Board.

Pennock-Román, M. (1994). College major and gender differences in the prediction of college grades (CBR 94-2). New York: The College Board.

Powers, D., & Dwyer, C. A. (2003). *Toward specifying a construct of reasoning* (RM-03-01). Princeton, NJ: Educational Testing Service.

Ragosta, M., Braun, H., & Kaplan, B. (1991). *Performance and persistence: A validity study of the SAT for students with disabilities* (College Board Report No. 91-3, ETS Research Report No. 91-41). New York: College Entrance Examination Board.

Ragosta, M., & Wendler, C. (1992). *Eligibility issues and comparable time limits for disabled and nondisabled SAT examinees* (College Board Research Report No. 92-5, ETS Research Report RR-92-35). New York: College Entrance Examination Board.

Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham et al. (Eds.), *Predicting college grades: An analysis of trends over two decades* (pp. 253–288). Princeton, NJ: Educational Testing Service.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (CBR 93-1). New York: The College Board.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin, 85*(6), 1348–1351.

Samejima, F. (1997). Graded response model. In Van Linden, W. J. & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Silver, E. A., Kilpatrick, J., & Schlesinger, B. (1990). Thinking through mathematics: Fostering inquiry and communication in mathematics classrooms. New York: The College Board.

Steen, L. A. (Ed.). (1997). Why numbers count: Quantitative literacy for tomorrow's America. New York: The College Board.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–617.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duign, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–375). New York: Springer-Verlag.

Swineford, F. (1974). *The test analysis manual* (SR-74-06). Princeton, NJ: Educational Testing Service.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 25, 2003, from www.education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

Thurstone, L. L. (1947). The calibration of test items. *American Psychologist*, *2*,103–104.

Walker, M. E. (2003). *Scaling issues associated with the SAT I: Writing Test* (Statistical Report SR-2003-12). Princeton, NJ: Educational Testing Service.

Walker, M. E. (2005). *Evaluation of decision tree items for March 2005 writing section scaling*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Walker, M. E., & Allspach, J. R. (2005). *Scaling the SAT writing section*. Presentation at College Board offices for College Board staff members, New York, NY.

Walker, M. E., Allspach, J., & Liu, J. (2004). *Scaling the new SAT® writing section: Finding the best solution* (Statistical Report 2004-61). Princeton, NJ: Educational Testing Service.

Walker, M. E., & Liu, J. (2003). *Scaling the new SAT writing test: Evidence from the 2003 field trial* (Statistical Report SR-2003-94). Princeton, NJ: Educational Testing Service.

Walker, M. E., & Liu, J. (2004). *Scaling issues associated with the new SAT writing test*. Paper presented at the annual meeting of the National Council on Measurement in Education, April 13–15, 2004, San Diego, CA.

Walker, M. E., Liu, J., & Allspach, J. R. (2005). *Scaling tests via nonlinear post-stratification methods.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Wang, X. B. (2006). Investigating the effect of new SAT test lengths on the performance of regular SAT examinees (CBRR 2006-9). New York: The College Board.

Webb, N. L. (2006a). *Alignment analysis of secondary language arts standards and the SAT Reasoning Test: Maine*. External report provided to the Maine Department of Education on April 10, 2006.

Webb, N. L. (2006b). *Alignment analysis of secondary mathematics standards and the SAT Reasoning Test: Maine*. External report provided to the Maine Department of Education on April 10, 2006.

Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades.* Princeton, NJ: Educational Testing Service.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Boston, MA: Allyn and Bacon.

Wilson, K. M. (1983). A review of research on the prediction of academic performance after the freshman year (CBRR 83-2). New York: The College Board.

Young, J. W. (2001). Differential Validity, differential prediction, and college admission testing: A comprehensive review and analysis (College Board Research Report No. 2001-6). New York: The College Board.

Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 289–301). New York: Routledge/Falmer.

Young, J. W., with Kobrin, J. L. (2001). Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis (CBRR 2001-6). New York: The College Board.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249.

APPENDICES

APPENDIX A—MAINE TECHNICAL ADVISORY COMMITTEE MEMBERS

Table A-1. 2013–14 MHSA: Maine Technical Advisory Committee Members

Member Name	*Member Affiliation*
Brian Gong	Executive Director, National Center for Improvement of Educational Assessment
Lenora Murray	Assistant Superintendent, MSAD #49
Stephen Slater	Assistant Director of Assessment, Oregon Department of Education
Betsy Webb	Superintendent of Schools, Bangor Public Schools
Martha Thurlow	Director, NCEO/University of Minnesota

Table A-2. 2013–14 MHSA: Science Bias and Sensitivity Committee Members

Name	*Department*
Lynne Adams	Augusta School Department (SPED)
Judy Carey	Catholic Charities (Blind/Visually Impaired)
Melvin Curtis	Retired (SPED)
Julia O’Brien-Merrill	Retired (ESL)
Rebecca Perez	Rumford Elementary School (ESL Teacher)

Table A-3. 2013–14 MHSA: Item Review Committee—Science

Name	*School*
Mary Whitten	Gardiner Area High School
Beth Chagrauslis	Brighton Academy
Lisa Damian-Marvin	Camden Hills Regional High School
Douglas Hodum	Mt. Blue High School
William Leathem	Hamden Academy
Patricia Spilecki	Lewiston High School
Amy Troiano	Westbrook High School
Sheree Granger	The School at Sweetser. Saco

APPENDIX B—POLICIES AND PROCEDURES

2013-2014 POLICIES AND PROCEDURES FOR ACCOMMODATIONS TO THE MAINE HIGH SCHOOL ASSESSMENT (MHSA)

The No Child Left Behind (NCLB) Act mandates that all students in one high school year be included in a state assessment. In addition, Maine Learning Results legislation requires that all students be included in a State assessment during their third year of high school. Maine's High School Accountability Assessment for students in their **third year of high school** consists of the SAT and a science test, which are administered at separate times.

Students will participate in these assessments through one of the following avenues: **Standard Administration, Administration with Accommodations, or Alternate Assessment (Personalized Alternate Assessment Portfolio [PAAP])**. Legal requirements for students identified for federally funded programs have been taken into account in the development of this document.

ACCOMMODATIONS

An accommodation removes a barrier that exists for a learner to allow access to the assessment without altering what is being measured. These policies and procedures for accommodations are designed so that all students with unique learning needs have a fair opportunity to demonstrate what they know and are able to do on all state required assessments at the high school level. All Maine students participating in state required assessments have access to the same accommodations, regardless of grade level.

The Maine High School Assessment provides two categories of accommodations:

1. **Maine Purposes Only (MPO),** approved only by a local team of educators which result in scores that measure a student's progress towards achievement of Maine's Learning Results for State and Federal purposes only.
2. **Services for Students with Disabilities (SSD),** approved by the College Board which result in scores that measure a student's progress towards achievement of Maine's Learning Results for State and Federal purposes **and** in SAT scores that can be used as part of a student's application for college admission.

DETERMINATION OF NEED FOR ACCOMMODATION

All students being considered for accommodations must have their individual situations reviewed by a team prior to the time of assessment. This team should include at least one of the student's teachers, the building principal, related services personnel, the parent(s)/guardian(s) and, whenever possible, the student. If it is not possible for the parent and student to attend the meeting, they should be consulted regarding the committee's recommendations for accommodations prior to the time of the assessment. The list of allowable accommodations than can be considered is located on pages 3-5 of this document.

- **Students without an Individual Educational Program (IEP) – MPO accommodations only**

  Students may include, but are not limited to, those who: are ill or incapacitated in some way; are Limited English Proficient (LEP); are identified as having disabilities under Section 504 of the Rehabilitation Act; or are identified by a team as needing accommodations in order to demonstrate an accurate level of academic achievement.

- **Students with an Individual Educational Program (IEP) - College-Board approved or MPO accommodations**

  Schools are required to address needed accommodations at an IEP Team meeting. Membership for this meeting is prescribed in Maine Unified Special Education Regulations, Chapter 101, July 19, 2013, which is located at: http://www.maine.gov/doe/specialed/laws/index.html. Only students with an identified disability under IDEA may be considered for accommodations for a standard SAT administration with resulting official College Board scores. If accommodations are either not submitted or not approved by the College Board, the students may use MPO accommodations but may not use their scores for college application purposes.

Procedures for Requesting College Board Approved Accommodations:

Students with an identified disability who need accommodations and wish to have college reportable scores on the SAT portion of their Maine High School Assessment or be eligible for scholarship programs through the PSAT/NMSQT, must file an official **College Board Eligibility Form**, identifying the accommodations they wish to use during the administration of the assessment in which they will participate. The accommodations for which a student may apply include:

- those listed by the College Board in the Eligibility Packet,
- those needed by individual students and allowed by the College Board but not listed in the Eligibility packet, and
- Maine accommodations listed on pages 3-5 of this document, approved through the College Board Eligibility Form in the "Other" category.

The required documentation must accompany the request for College Board approved accommodations. The College Board will determine whether the use of the accommodations requested will be approved for the use of the individual student, based on their review.

DOCUMENTATION OF ACCOMMODATIONS

Any accommodations approved for a student and the reasons for these choices must be documented in a statement in the student's cumulative folder or in the IEP for a student with an identified disability. Refer to pages 3-5 of this document for the allowable accommodation codes for the Maine High School Assessment when taken for Maine Purposes Only.

ADMINISTRATION OF ACCOMMODATIONS

Test Center (School) personnel should be familiar with and administer all allowed accommodations in accordance with the directions provided in trainings for SAT Test Site Supervisors and those included in the Maine High School Assessment Administrators' Manual. The same accommodations must be provided for all components of the Maine High School Assessment. Coding of Maine Purposes Only accommodations (see pages 3-5 of this document) to be used by individual students will be entered by school personnel according to the directions provided by the College Board.

REPORTING STUDENTS' SCORES

Official SAT Reports

Free official SAT score reports will be issued to three colleges identified by a student who took the SAT portion of the Maine High School Assessment **with accommodations approved by the College Board**. The student will receive the report within 2 months of taking the SAT. Students using MPO accommodations for the MHSA will not get an official College Board score report.

Maine Reports for All Students

All students taking the Maine High School Assessment will be included in the school's accountability system and their scores will be included in the State assessment reports. The scores on these reports will be determined by the combination of the SAT and the Science component based on Maine's achievement standards and will be provided to schools at the beginning of the school year following testing.

Remember: Scores for students who use MPO accommodations on the SAT portion of the test cannot be sent to colleges by the College Board.

Approved Maine Purposes Only (MPO) Accommodations for the MHSA

Use of these accommodations without College Board approval through the Eligibility Process will result in scores reportable for Maine Purposes Only. All accommodations used must: not change what is being measured, be approved for individual students by a team, and be a regular part of the student's daily instruction.

Code	**Accommodations Category**	**Details on Delivery of Accommodations**
T	**TIMING –** Tests were administered:	
MT1	with time extended beyond standard administration (same day).	Extended time may be needed by students who are unable to meet time constraints, are easily fatigued, or unable to concentrate for the length of time allotted for test completion. Testing may be extended until student can no longer sustain the activity.
MT2	with time extended beyond standard administration (several days).	
MT3	with multiple or frequent breaks.	Multiple or frequent breaks may be required by students whose attention span, distractibility, or physical condition, require shorter working periods.
MT4	at a time of day or a day of the week most beneficial to the student.	Individual scheduling may be used for students whose school performance is noticeably affected by the time of day or day of the school week on which it is done.
MT5	using flexibility in the order in which content area tests are given.	Flexibility in the order of presentation may be used, for example, to build confidence in the student by testing those content areas in which they are strongest first, or to alleviate concerns by allowing them to complete the content area about which they are most apprehensive first.
S	**SETTING –** Tests were administered:	
MS1	in school site other than regular classroom.	Students may be tested in an alternative site to reduce distractions for themselves or others, or to increase physical access to special equipment.
MS2	in out-of-school setting by school personnel.	Out-of-school testing may be used for students who are hospitalized or unable to attend school.

Code	Accommodations Category	Details on Delivery of Accommodations
P	**PRESENTATION –** Tests were administered:	
MP1	individually.	Individual or small group testing may be used to minimize distractions for students whose test is administered out of the classroom or so that others will not be distracted by accommodations being used (ex., dictation).
MP2	in a small group.	
MP3	using a human reader.	A human reader may be used for a student whose inability to read would hinder performance. A Reader's Script will be provided based on registration with this accommodation. **NOTE:** When used for the Reading Passages, MP3 becomes a modification that is not allowed on other State assessments.
MP4	using sign language (*NOT allowed for reading passages).*	Trained personnel may use sign language to administer the test for deaf or hearing impaired students, with the exception of the reading passages. Sign language may be used only for questions and directions in the reading sessions.
MP5	with opportunity for student to move, stand, and/or pace during assessment.	This opportunity may be used in a single-student setting other than the classroom for a student who cannot focus when seated for sustained periods of time.
MP6	using alternative or assistive technology that is part of the student's communication system.	The test may be presented through his/her regular communication system to a student who uses alternative and assistive technology on a daily basis.
MP7	by school personnel known to the student other than the student's classroom teacher (e.g., ESL Title I, Special Education)	The test administrator may be a member of the staff who works with the student from time-to-time or on a daily basis, but is not the student's regular teacher for general curriculum.
MP8	using large print version of assessment.	A 20 pt. photo-enlarged print version of the SAT will be supplied based on registration with this accommodation.
MP9	using Braille version of assessment.	A Braille version of the SAT will be supplied based on registration with this accommodation.
MP10	with LEP student use of a word-to-word bilingual dictionary as needed.	The student may have a word for word dictionary available for individual use as needed. A word for word dictionary is one that does not include any definitions. Dictionaries used must be among those listed at http://www.maine.gov/doe/mhsa/administration/index.html.
MP12	using a cassette version of the test.	A cassette version of the SAT will be supplied based on registration with this accommodation.

Code	Accommodations Category	Details on Delivery of Accommodations
R	**RESPONSE –** Tests were administered:	
MR1	using a scribe or recording device *(oral dictation to a scribe or a recording device is NOT allowed for the Writing session).*	The student may dictate answers to trained personnel or record answers in an individual setting so that other students will not benefit by hearing answers or be otherwise disturbed. Recorded answers must be scribed prior to the return of test materials. Audio recordings must be deleted immediately following scribing.
MR2	using alternative or assistive technology/devices that are part of the student's communication system.	The technology is used to permit the student to read and/or respond to the test. In addition to computers, such devices might include, for example, text enlargers, speech-to-text, amplification devices, Dynaboxes, etc. Speech-to-text may not be used for the Writing session.
MR3	other assistive devices.	To enable a student to organize thinking, focus, and/or use a device that serves as a specific strategy related to a test item, other assistive devices may be used. They might include such things as templates, graphic organizers, arithmetic tables *(only in the calculator allowed session of the Mathematics test)*, noise buffers, place markers, carrels, etc.
MR4	with student use of a word processor. **MHSA ONLY**	A student may use a word processor. When used for the Writing session, spell check, grammar check, and word prediction programs should be turned off.
MR5	with student use of a Brailler. **MHSA ONLY**	A student may use a Braillewriter, a slate and stylus, and/or an electronic Brailler to respond to questions. Responses would need to be recorded in standard format by a scribe.
MR6	with student use of visual aids.	Visual aids include any optical or non-optical devices used to enhance visual capability. Examples include magnifiers, special lighting, markers, filters, large-spaced paper, color overlays, etc.
MR7	with LEP student use of a word-to-word bilingual dictionary as needed.	The student may have a word for word dictionary available for individual use as needed. A word for word dictionary is one that does not include any definitions. Dictionaries used must be among those listed at http://www.maine.gov/doe/mhsa/administration/index.html
MR8	using administrator verification of student understanding following the reading of test directions.	After directions have been read, the test administrator may ask the student what he/she has been asked to do. If directions have been misunderstood by the student, the directions may be paraphrased or demonstrated. Test items may not be paraphrased or explained.
MR9	using side-by-side placement of two test booklets.	All responses must be recorded on a single answer sheet. This accommodation is designed to allow students to see all sections related to the same item at the same time, regardless of the test configuration.
O	**Other**	Must be documented and submitted to the Department of Education in advance. **Contact:** **Susan Fossett**, Assessment Coordinator susan.fossett@maine.gov; 207- 624-6775

APPENDIX C—PARTICIPATION RATES SCIENCE

Table C-1. 2013–14 MHSA: Summary of Participation by Demographic Category—Science

Description	*Tested*	
	Number	*Percent*
All Students	12,761	100.00
Male	6,594	51.67
Female	6,167	48.33
Not Reported	0	0.00
Hispanic or Latino	185	1.45
American Indian or Alaskan Native	92	0.72
Asian	165	1.29
Black or African American	405	3.17
Native Hawaiian or Pacific Islander	14	0.11
White (non-Hispanic)	11,774	92.27
Two or more races	126	0.99
Currently LEP student	247	1.94
Former LEP student – monitoring year 1	35	0.27
Former LEP student – monitoring year 2	59	0.46
All Other Students	12,420	97.33
Students with an IEP	1,679	13.16
All Other Students	11,082	86.84
Economically Disadvantaged Students	4,581	35.90
All Other Students	8,180	64.10
Migrant Students	2	0.02
All Other Students	12,759	99.98
Students Receiving Title 1 Services	227	1.78
All Other Students	12,534	98.22
Students with a 504 plan	590	4.62
All Other Students	12,171	95.38

Appendix D—National Tables

Table D-1. 2013–14 MHSA: National Summary Statistics for the Critical Reading Test of the College Board SAT

Form	*1*	*2*
Administration	*May 2014*	*June 2014*
Total Group Statistics[a]		
Total Group N	411,679	426,840
Formula Score Information		
Mean	33.1	33.2
SD	14.6	15.0
Possible range	-17-67	-17-67
Obtained range	-12-67	-13-67
Median	34	34
Skewness	-.13	-.11
Scaled Score Information		
Mean	504	506
SD	106	105
Possible range[b]	200-860	200-880
Obtained range[b]	200-860	200-880
Median	500	510
Sample Statistics[c]		
Sample N	15,048	10,392
Formula Score Information		
Mean	34.3	33.8
SD	14.2	14.8
Obtained range	-8-67	-7-67
Median	35	35
Skewness	-.18	-.16
Scaled Score Information		
Mean	513	510
SD	104	104
Obtained range[b]	200-860	200-880
Median	510	510
Item Information		
Number of items	67	67
Mean proportion correct	.59	.58
Mean observed delta	11.8	11.8
SD observed delta	2.5	2.5
Mean equated delta	11.4	11.4
SD equated delta	2.3	2.3
Mean *r*-biserial	.53	.55
SD *r*-biserial	.09	.11
No. *r*-biserial < 0.20	0	0

[a] Total group statistics are based on all on-time cases, regardless of grade levels.
[b] If scores are not truncated at 800 and not extrapolated below 200
[c] Sample statistics are based on a spaced random sample of juniors and seniors and are reported to be directly comparable to the “target,” or typical, population of test takers—juniors or seniors expected to soon go to college.
SD = standard deviation

Table D-2. 2013–14 MHSA: National Summary Statistics for the Mathematics Test of the College Board SAT*

Form *Administration*	*1* May 2014	*2* June 2014
Total Group Statistics[a]		
Total Group N	411,679	426,840
Formula Score Information		
Mean	27.9	28.1
SD	12.8	13.1
Possible range	-11-54	-11-54
Obtained range	-8-54	-9-54
Median	28	28
Skewness	-.07	-.04
Scaled Score Information		
Mean	514	513
SD	107	109
Possible range[b]	200-810	200-800
Obtained range[b]	200-810	200-800
Median	510	510
Sample Statistics[c]		
Sample N	15,048	10,392
Formula Score Information		
Mean	28.4	28.8
SD	12.7	13.0
Obtained range	-6-54	-5-54
Median	29	29
Skewness	-.11	-.08
Scaled Score Information		
Mean	518	519
SD	107	108
Obtained range[b]	200-810	200-800
Median	520	520
Item Information		
Number of items	54	54
Mean proportion correct	0.58	0.58
Multiple-Choice Items		
Mean observed delta	11.6	11.5
SD observed delta	2.6	2.5
Mean equated delta	12.2	12.1
SD equated delta	3.0	3.0
Mean *r*-biserial	.61	.60
SD *r*-biserial	.10	.12
No. *r*-biserial < 0.20	0	0
Student-Produced-Response Items		
Mean observed delta	13.0	12.5
SD observed delta	2.1	2.1
Mean equated delta	13.8	13.3
SD equated delta	2.4	2.5
Mean *r*-biserial	.68	.74
SD *r*-biserial	.08	.04
No. *r*-biserial < 0.20	0	0

[a] Total group statistics are based on all on-time cases, regardless of grade levels.

[b] If scores are not truncated at 800 and not extrapolated below 200

[c] Sample statistics are based on a spaced random sample of juniors and seniors and are reported to be directly comparable to the "target," or typical, population of test takers—juniors or seniors expected to soon go to college.

SD = standard deviation

Table D-3. 2013–14 MHSA: National Summary Statistics for the Writing Test of the College Board SAT

Form *Administration*	*1* May 2014	*2* June 2014
Total Group Statistics[a]		
Total Group N	411,679	426,840
Formula Score Information		
Mean	24.9	26.6
SD	10.5	10.4
Possible range	-12-49	-12-49
Obtained range	-10-49	-8-49
Median	25	27
Skewness	.04	-.11
Scaled Score Information		
Mean	487	500
SD	105	107
Possible range[b]	200-800	200-810
Obtained range[b]	200-800	200-810
Median	480	500
Sample Statistics[c]		
Sample N	15,048	10,392
Formula Score Information		
Mean	25.6	27.1
SD	10.3	10.3
Obtained range	-7-49	-6-49
Median	25	27
Skewness	.01	-.14
Scaled Score Information		
Mean	493	505
SD	103	106
Obtained range[b]	200-800	200-810
Median	480	500
Item Information		
Number of items	49	49
Mean proportion correct	.61	.63
Mean observed delta	11.6	11.3
SD observed delta	2.6	2.7
Mean equated delta	10.1	10.1
SD equated delta	2.4	2.5
Mean *r*-biserial	.54	.55
SD *r*-biserial	.10	.09
No. *r*-biserial < 0.20	0	0

[a] Total group statistics are based on all on-time cases, regardless of grade levels.
[b] If scores are not truncated at 800 and not extrapolated below 200
[c] Sample statistics are based on a spaced random sample of juniors and seniors.
SD = standard deviation

Table D-4. 2013–14 MHSA: National Raw to Scaled Score Conversion Table for the Critical Reading Test of the College Board SAT

Form	*1*	*2*	*Form*	*1*	*2*
Administration	*May 2014*	*June 2014*	*Administration*	*May 2014*	*June 2014*
Raw Score	*Rounded Scaled Score*		*Raw Score*	*Rounded Scaled Score*	
67	800	800	24	440	450
66	800	800	23	440	440
65	800	800	22	430	430
64	790	800	21	420	430
63	770	780	20	420	420
62	760	760	19	410	420
61	740	740	18	400	410
60	730	720	17	400	400
59	720	710	16	390	400
58	700	700	15	380	390
57	690	690	14	380	390
56	680	670	13	370	380
55	670	660	12	360	370
54	660	650	11	360	360
53	650	640	10	350	360
52	640	630	9	340	350
51	630	630	8	330	340
50	620	620	7	320	330
49	620	610	6	310	320
48	610	600	5	300	310
47	600	590	4	290	300
46	590	590	3	270	290
45	580	580	2	250	270
44	580	570	1	240	260
43	570	560	0	220	240
42	560	560	-1	200	220
41	550	550	-2	200	200
40	550	540	-3	200	200
39	540	540	-4	200	200
38	530	530	-5	200	200
37	520	520	-6	200	200
36	520	520	-7	200	200
35	510	510	-8	200	200
34	500	510	-9	200	200
33	500	500	-10	200	200
32	490	490	-11	200	200
31	480	490	-12	200	200
30	480	480	-13	200	200
29	470	480	-14	200	200
28	470	470	-15	200	200
27	460	460	-16	200	200
26	450	460	-17	200	200
25	450	450			

Table D-5. 2013–14 MHSA: National Raw to Scaled Score Conversion Table for the Mathematics Test of the College Board SAT

Form	*1*	*2*	*Form*	*1*	*2*
Administration	*May 2014*	*June 2014*	*Administration*	*May 2014*	*June 2014*
Raw Score	*Rounded Scaled Scores*		*Raw Score*	*Rounded Scaled Scores*	
54	800	800	21	460	460
53	780	770	20	450	450
52	750	740	19	450	440
51	730	720	18	440	440
50	710	700	17	430	430
49	700	690	16	420	420
48	690	680	15	420	410
47	670	670	14	410	410
46	660	660	13	400	400
45	650	650	12	390	390
44	640	640	11	380	380
43	630	630	10	370	370
42	620	620	9	360	360
41	620	610	8	360	350
40	610	600	7	350	340
39	600	590	6	330	330
38	590	580	5	320	320
37	580	580	4	310	310
36	570	570	3	300	290
35	560	560	2	280	280
34	560	550	1	270	260
33	550	550	0	250	250
32	540	540	-1	230	230
31	530	530	-2	210	210
30	530	520	-3	200	200
29	520	520	-4	200	200
28	510	510	-5	200	200
27	500	500	-6	200	200
26	500	490	-7	200	200
25	490	490	-8	200	200
24	480	480	-9	200	200
23	470	470	-10	200	200
22	470	470	-11	200	200

Table D-6. 2013–14 MHSA: National Raw to Scaled Score Conversion Table for the Multiple-Choice Score of the Writing Test of the College Board SAT

Form	*1*	*2*	*Form*	*1*	*2*
Administration	*May 2014*	*June 2014*	*Administration*	*May 2014*	*June 2014*
Raw Score	*Rounded Scaled Scores*		*Raw Score*	*Rounded Scaled Scores*	
49	800	800	18	420	410
48	770	780	17	410	400
47	750	750	16	400	390
46	730	730	15	390	390
45	710	710	14	390	380
44	690	690	13	380	370
43	680	670	12	370	360
42	660	660	11	360	350
41	650	650	10	350	340
40	640	630	9	340	330
39	630	620	8	330	320
38	620	610	7	320	310
37	600	600	6	310	300
36	590	590	5	290	280
35	580	580	4	280	270
34	570	570	3	260	250
33	560	560	2	250	240
32	550	550	1	230	220
31	540	540	0	210	200
30	530	530	-1	200	200
29	520	520	-2	200	200
28	510	510	-3	200	200
27	500	500	-4	200	200
26	490	490	-5	200	200
25	480	480	-6	200	200
24	470	470	-7	200	200
23	460	460	-8	200	200
22	450	450	-9	200	200
21	450	440	-10	200	200
20	440	430	-11	200	200
19	430	420	-12	200	200

Table D-7. 2013–14 MHSA: Reliability Coefficients and Standard Errors of Measurement[a] for Sections of the College Board SAT—National Equating Sample

		Form	*1*		*2*	
		Administration	May 2014		June 2014	
		Sample N	15,048		10,392	
Test Section			*Rel.*	*SEM*	*Rel.*	*SEM*
Critical reading 1	Dressel-KR20		.80	*2.4*	.81	*2.4*
Critical reading 2	Dressel-KR20		.81	*2.5*	.83	*2.4*
Critical reading 3	Dressel-KR20		.76	*2.2*	.78	*2.2*
Total critical	Kristof	Raw	.90	*4.4*	.92	*4.3*
Reading	Var. components	Raw	.92	*4.1*	.93	*4.0*
	IRT[b]	Raw	.92	*4.1*	.93	*4.1*
	IRT[b]	Scaled	.91	*31*	.92	*30*
Math 1	Dressel-KR20		.82	*2.1*	.82	*2.1*
Math 2	Dressel-KR20		.85	*1.8*	.86	*1.7*
Math 3	Dressel-KR20		.80	*1.8*	.81	*1.9*
Total mathematics	Kristof	Raw	.93	*3.3*	.93	*3.3*
	Var. components	Raw	.93	*3.3*	.94	*3.3*
	IRT[b]	Raw	.93	*3.4*	.93	*3.4*
	IRT[b]	Scaled	.93	*29*	.93	*29*
Writing 1	Dressel-KR20		.84	*3.0*	.85	*2.9*
Writing 2	Dressel-KR20		.73	*1.8*	.72	*1.7*
Total writing MC	Angoff/Feldt	Raw	.87	*3.7*	.88	*3.5*
	Var. components	Raw	.88	*3.5*	.89	*3.4*
	IRT[b]	Raw	.89	*3.5*	.89	*3.4*
	IRT[b]	Scaled	.89	*35*	.89	*35*
Writing composite[c]		Scaled	.89	*34*	.90	*33*

[a] See Appendix H for formulas used to compute reliability coefficients and SEM.

[b] See SR-84-118 for a description of algorithms employed for IRT based statistics.

[c] See “Computing Easy SAT Writing Section Score Reliability Estimates” by Michael E. Walker, issued February 13, 2006, for a description of the methods used in calculating reliability and SEM for the writing composite scores.

Table D-8. 2013–14 MHSA: National Completion Rate Statistics for Sections of the College Board SAT

Form	*1*	*2*	*1*	*2*	*1*	*2*
Administration	5/14	6/14	5/14	6/14	5/14	6/14
Sample N	15,048	10,392	15,048	10,392	15,048	10,392
	Critical Reading 1		Mathematics 1		Writing 1	
% completing section	87.1	80.1	57.4	51.0	75.3	74.2
% completing 75%	99.8	99.4	97.2	98.5	100.0	100.0
No. items reached by 80%	23	24	18	19	34	34
Mean not reached	0.2	0.6	1.0	0.8	0.6	0.6
SD not reached	0.8	1.5	1.6	1.2	1.2	1.2
NR variance/score variance	0.02	0.07	0.10	0.06	0.03	0.02
No. items in section	23	24	20	20	35	35
	Critical Reading 2		Mathematics 2		Writing 2	
% completing section	84.8	81.3	44.2	48.9	88.7	92.8
% completing 75%	97.8	98.3	96.7	94.2	99.2	99.6
No. items reached by 80%	25	24	15	16	14	14
Mean not reached	0.6	0.6	1.3	1.1	0.2	0.1
SD not reached	1.7	1.6	1.6	1.7	0.7	0.6
NR variance/score variance	0.09	0.08	0.13	0.13	0.04	0.03
No. items in section	25	24	18	18	14	14
	Critical Reading 3		Mathematics 3			
% completing section	75.3	84.5	69.1	54.6		
% completing 75%	97.4	97.8	98.4	96.1		
No. items reached by 80%	18	19	15	14		
Mean not reached	0.5	0.4	0.5	1.0		
SD not reached	1.3	1.3	1.1	1.5		
NR variance/score variance	0.08	0.08	0.07	.12		
No. items in section	19	19	16	16		

Table D-9. 2013–14 MHSA: National Summary Statistics of Equated Deltas (Δ) for Critical Reading, Mathematics, and Writing Sections of the College Board SAT

Form		*Specified*	*1*		*2*	
Administration			5/14		6/14	
Sample N			15,048		10,392	
		Specified Equated Delta	Equated Delta	Observed Delta	Equated Delta	Observed Delta
Total Critical Reading	N	67	67	67	67	67
	Mean	11.4	11.4	11.8	11.4	11.8
	SD	2.4	2.3	2.5	2.3	2.5
Mathematics MC	N	44	44	44	44	44
	Mean	12.2	12.2	11.6	12.1	11.5
	SD	3.2	3.0	2.6	3.0	2.5
Mathematics SPR	N	10	10	10	10	10
	Mean	13.6-14.2	13.8	13.0	13.3	12.5
	SD	3.0	2.4	2.1	2.5	2.1
Total Writing	N	49	49	49	49	49
	Mean	10.1	10.1	11.6	10.1	11.3
	SD	2.5	2.4	2.6	2.5	2.7

MC = multiple-choice; SPR = student-produced-response; SD = standard deviation

Table D-10. 2013–14 MHSA: National Summary Statistics for Biserial Coefficients for Critical Reading, Mathematics, and Writing Sections of the College Board SAT

	Form	*Specified*	*1*	*2*
	Administration		5/14	6/14
	Sample N		15,048	10,392
Total critical reading	N		67	67
	Not comp.[a]		0	0
	Mean	0.49-0.53[b]	0.53	0.55
	SD		0.09	0.11
Mathematics MC	N		44	44
	Not comp.[a]		0	0
	Mean	0.53-0.57[b]	0.61	0.60
	SD		0.10	0.12
Mathematics SPR	N		10	10
	Not comp.[a]		0	0
	Mean	0.60-0.70[b]	0.68	0.74
	SD		0.08	0.04
Total writing	N		49	49
	Not comp.[a]		0	0
	Mean	0.49-0.53[b]	0.54	0.55
	SD		0.10	0.09

[a] *R*-biserial is not calculated when the percentage correct is greater than 95 or less than 5, or when dropout exceeds 50%.

[b] Mean r-biserial is specified in terms of final-form items which are included in the criterion. The equivalent mean for a total-score criterion that does not include the item is 0.45 – 0.49 for total critical reading.

MC = multiple-choice; SPR = student-produced-response; SD = standard deviation

Table D-11. 2013–14 MHSA: National Differential Item Functioning (DIF) Summary Form—May 2014 Administration

Category of Maximum Absolute DIF Value for All Comparisons			*Female*	*Black*	*Hispanic*	*Asian*	*Am. Indian*
			N=188,367	N=43,347	N=47,805	N=24,934	N=2,065
			Male	White	White	White	White
			N=164,159	N=216,844	N=216,844	N=216,844	N=216,844
Category	Number	% of Items	Number of Items by DIF Category				
Total critical reading							
+C	0	0.0	0	0	0	0	0
+B	1	1.5	0	0	1	0	0
A	62	92.5	64	67	66	66	67
-B	4	6.0	3	0	0	1	0
-C	0	0.0	0	0	0	0	0
Total	67	100.0	67	67	67	67	67
Total mathematics							
+C	0	0.0	0	0	0	0	0
+B	3	5.6	0	1	0	2	0
A	49	90.7	52	53	54	51	54
-B	2	3.7	2	0	0	1	0
-C	0	0.0	0	0	0	0	0
Total	54	100.0	54	54	54	54	54
Total writing							
+C	0	0.0	0	0	0	0	0
+B	3	6.1	1	0	0	2	0
A	42	85.7	46	49	49	44	49
-B	4	8.2	2	0	0	3	0
-C	0	0.0	0	0	0	0	0
Total	49	100.0	49	49	49	49	49

Table D-12. 2013–14 MHSA: National Differential Item Functioning (DIF) Summary Form—June 2014 Administration

Category of Maximum Absolute DIF Value for All Comparisons			*Female*	*Black*	*Hispanic*	*Asian*	*Am. Indian*
			N=202,514	N=46,768	N=57,626	N=32,045	N=2,029
			Male	*White*	*White*	*White*	*White*
			N=159,133	N=204,510	N=204,510	N=204,510	N=204,510
Category	*Number*	*% of Items*	*Number of Items by DIF Category*				
Total critical reading							
+C	1	1.5	1	0	0	0	0
+B	2	3.0	1	0	1	0	0
A	59	88.1	62	65	63	66	67
-B	3	4.5	2	2	2	1	0
-C	2	3.0	1	0	1	0	0
Total	67	100.0	67	67	67	67	67
Total mathematics							
+C	0	0.0	0	0	0	0	0
+B	0	0.0	0	0	0	0	0
A	47	87.0	48	50	54	53	54
-B	7	13.0	6	4	0	1	0
-C	0	0.0	0	0	0	0	0
Total	54	100.0	54	54	54	54	54
Total writing							
+C	0	0.0	0	0	0	0	0
+B	3	6.1	0	1	1	3	0
A	44	89.8	48	48	48	44	49
-B	2	4.1	1	0	0	2	0
-C	0	0.0	0	0	0	0	0
Total	49	100.0	49	49	49	49	49

APPENDIX E—ITEM-LEVEL CLASSICAL STATISTICS SCIENCE

Table E-1. 2013–14 MHSA: Item-Level Classical Test Theory Statistics—Science Grade 11

Item		*Difficulty*	*Discrimination*	*Percent Omitted*
Number	*Type*			
466	MC	0.68	0.20	2
2924	MC	0.45	0.34	6
68925	MC	0.57	0.34	2
68972	MC	0.78	0.47	1
74219	MC	0.70	0.33	2
96884	MC	0.55	0.40	6
96888	MC	0.54	0.40	2
97943	MC	0.69	0.36	1
98063	MC	0.42	0.31	7
142385	MC	0.48	0.38	3
153419	MC	0.74	0.33	1
154096	MC	0.72	0.47	2
187043	MC	0.41	0.38	3
187222	MC	0.49	0.21	2
187231	MC	0.66	0.42	3
187232	MC	0.29	0.20	5
187244	MC	0.75	0.28	1
187263	MC	0.45	0.17	9
228231	MC	0.67	0.40	3
228233	MC	0.79	0.29	2
228236	MC	0.46	0.40	6
228242	MC	0.57	0.36	2
228259	MC	0.42	0.29	6
228275	MC	0.58	0.36	2
228283	MC	0.46	0.30	9
228286	MC	0.88	0.25	1
228288	MC	0.72	0.38	3
228289	MC	0.78	0.39	1
228313	MC	0.49	0.31	3
228320	MC	0.25	0.32	3
236887	CR	0.47	0.56	6
248417	MC	0.57	0.31	5
248418	MC	0.62	0.31	5
248421	MC	0.55	0.33	5
248428	MC	0.49	0.17	4
248430	MC	0.46	0.19	2
248437	CR	0.32	0.64	14
248439	CR	0.31	0.36	5
248450	MC	0.47	0.24	6
248452	MC	0.36	0.27	17
248457	MC	0.61	0.32	8
248465	MC	0.33	0.25	7
248468	MC	0.54	0.37	6
248504	CR	0.31	0.53	7

APPENDIX F—ITEM-LEVEL SCORE POINT DISTRIBUTIONS SCIENCE

Table F-1. 2013–14 MHSA: Item-Level Score Point Distributions for Constructed Response Items—Science

Subject	*Grade*	*Item Number*	*Total Possible Points*	*Percent of Students at Each Score Point*				
				0	*1*	*2*	*3*	*4*
Science	11	236887	4	7.71	23.76	30.46	23.75	8.50
		248437	4	23.71	23.61	19.29	13.74	6.14
		248439	4	30.60	23.61	26.49	9.88	3.97
		248504	4	19.61	37.88	24.22	9.61	1.93

APPENDIX G—DIFFERENTIAL ITEM FUNCTIONING RESULTS SCIENCE

Table G-1. 2013–14 MHSA: Number of Items Classified as “Low” or “High” DIF Overall and by Group Favored—Science

Grade	*Group*		*Item Type*	*Number of Items*	*Number “low”*			*Number “high”*		
	Reference	*Focal*			*Total*	*Favoring*		*Total*	*Favoring*	
						Reference	*Focal*		*Reference*	*Focal*
11	Male	Female	MC	40	9	7	2	4	4	0
			OR	4	2	0	2	1	0	1
	White	Black	MC	40	7	5	2	1	1	0
			OR	4	0	0	0	0	0	0
	Non-EconDis	EconDis	MC	40	0	0	0	0	0	0
			OR	4	0	0	0	0	0	0
	No Disability	Disability	MC	40	6	5	1	2	0	2
			OR	4	1	1	0	2	2	0
	Non-LEP	LEP	MC	40	10	8	2	3	2	1
			OR	4	0	0	0	0	0	0

EconDis = economically disadvantaged; LEP = limited English proficient.
MC = multiple-choice; OR = open-response.

APPENDIX H—ITEM RESPONSE THEORY PARAMETERS SCIENCE

Table H-1. 2013–14 MHSA: Item Response Theory Parameters— Science

Item		*a*	*SE (a)*	*b*	*SE (b)*	*D1*	*SE (D1)*	*D2*	*SE (D2)*	*D3*	*SE (D3)*	*D4*	*SE (D4)*	*D5*	*SE (D5)*
Number	*Type*														
153419	MC	0.88755	0	-0.70409	0	0.06180	0	-0.06180	0	0	0				
228289	MC	0.74403	0	-0.91401	0	0.02735	0	-0.02735	0	0	0				
187231	MC	0.75237	0	-0.45179	0	0.08740	0	-0.08740	0	0	0				
228286	MC	0.21209	0	0	0	0.06164	0	-0.06164	0	0	0				
248417	MC	0.54694	0	-0.32890	0	0.11550	0	-0.11550	0	0	0				
154096	MC	0.95165	0	-0.46184	0	0.03521	0	-0.03521	0	0	0				
228242	MC	0.47384	0	-0.36220	0	0.06086	0	-0.06086	0	0	0				
228283	MC	0.50994	0	0.17977	0	0.21382	0	-0.21382	0	0	0				
228313	MC	0.49850	0	0.07506	0	0.09477	0	-0.09477	0	0	0				
466	MC	0.32980	0	-1.41722	0	0.14449	0	-0.14449	0	0	0				
142385	MC	0.65493	0	0.30676	0	0.08434	0	-0.08434	0	0	0				
248421	MC	0.47927	0	-0.26051	0	0.11305	0	-0.11305	0	0	0				
98063	MC	0.53164	0	0.71234	0	0.17002	0	-0.17002	0	0	0				
248428	MC	0.50719	0	-0.73885	0	0.09124	0	-0.09124	0	0	0				
97943	MC	0.54222	0	-0.69639	0	0.04006	0	-0.04006	0	0	0				
96888	MC	0.65630	0	-0.28233	0	0.06733	0	-0.06733	0	0	0				
187222	MC	0.43063	0	0.12604	0	0.08319	0	-0.08319	0	0	0				
187244	MC	0.50560	0	-0.81255	0	0.04836	0	-0.04836	0	0	0				
228288	MC	0.85419	0	-0.86352	0	0.07487	0	-0.07487	0	0	0				
187043	MC	0.67824	0	0.30211	0	0.05553	0	-0.05553	0	0	0				
68972	MC	0.87382	0	-0.87725	0	0.03147	0	-0.03147	0	0	0				
228320	MC	0.62544	0	1.75059	0	0.11190	0	-0.11190	0	0	0				
248457	MC	0.40967	0	-0.50099	0	0.28385	0	-0.28385	0	0	0				
248430	MC	0.26223	0	0.69971	0	0.08856	0	-0.08856	0	0	0				
248450	MC	0.34432	0	-0.25948	0	0.16561	0	-0.16561	0	0	0				
228259	MC	0.39740	0	0.58289	0	0.19787	0	-0.19787	0	0	0				
248418	MC	0.49141	0	-0.50685	0	0.16174	0	-0.16174	0	0	0				
68925	MC	0.53542	0	0.08421	0	0.06050	0	-0.06050	0	0	0				
187232	MC	0.32137	0	2.07467	0	0.25648	0	-0.25648	0	0	0				
248465	MC	0.33634	0	1.22456	0	0.33564	0	-0.33564	0	0	0				
96884	MC	0.61565	0	-0.15753	0	0.18397	0	-0.18397	0	0	0				

continued

Item		*a*	*SE (a)*	*b*	*SE (b)*	*D1*	*SE (D1)*	*D2*	*SE (D2)*	*D3*	*SE (D3)*	*D4*	*SE (D4)*	*D5*	*SE (D5)*
Number	*Type*														
228275	MC	0.55671	0	-0.34705	0	0.10411	0	-0.10411	0	0	0				
248468	MC	0.48150	0	0.30973	0	0.10864	0	-0.10864	0	0	0				
228233	MC	0.57545	0	-1.50060	0	0.06336	0	-0.06336	0	0	0				
248452	MC	0.31155	0	0.30380	0	0.76180	0	-0.76180	0	0	0				
74219	MC	0.47723	0	-0.66905	0	0.07957	0	-0.07957	0	0	0				
2924	MC	0.47736	0	0.78269	0	0.24178	0	-0.24178	0	0	0				
187263	MC	0.19166	0	-0.72487	0	0.60733	0	-0.60733	0	0	0				
228231	MC	0.60218	0	-1.02877	0	0.07851	0	-0.07851	0	0	0				
228236	MC	0.54615	0	0.15145	0	0.11779	0	-0.11779	0	0	0				
248504	CR	0.78646	0	2.10439	0	2.13520	0	0.74021	0	-0.45407	0	-2.42134	0	0	0
248437	CR	1.06923	0	0.78251	0	1.34884	0	0.44738	0	-0.44865	0	-1.34758	0	0	0
248439	CR	0.55825	0	1.12225	0	2.47345	0	0.94452	0	-0.91708	0	-2.50089	0	0	0
236887	CR	0.70402	0	0.83091	0	2.54567	0	1.05424	0	-0.68693	0	-2.91298	0	0	0

* Note that these items were pre-equated so the standard error estimation is not reflecting the current year data.

APPENDIX I—TEST CHARACTERISTIC CURVES AND TEST INFORMATION FUNCTIONS
SCIENCE

Figure I-1. 2013–14 MHSA: Test Characteristic Curve

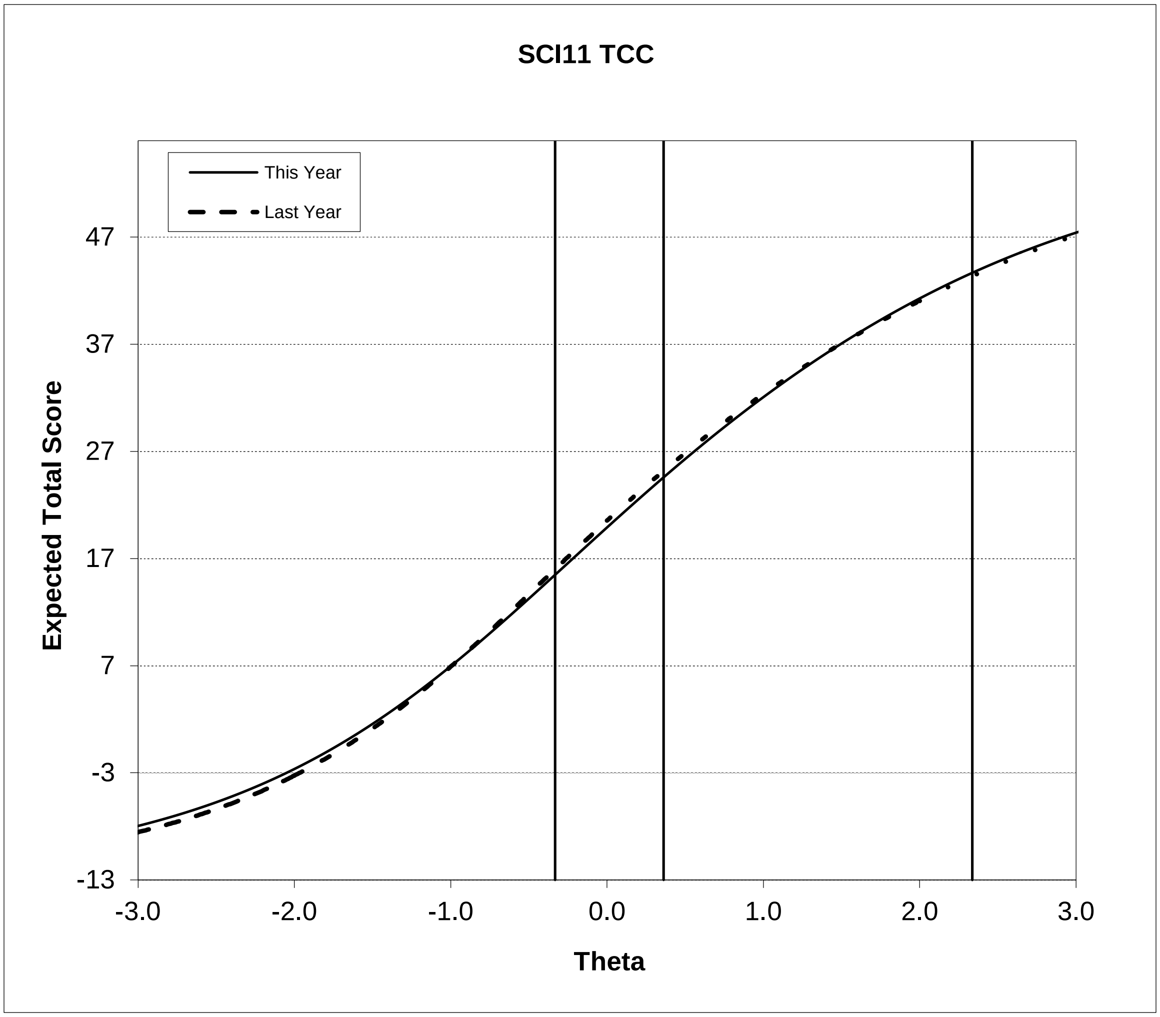
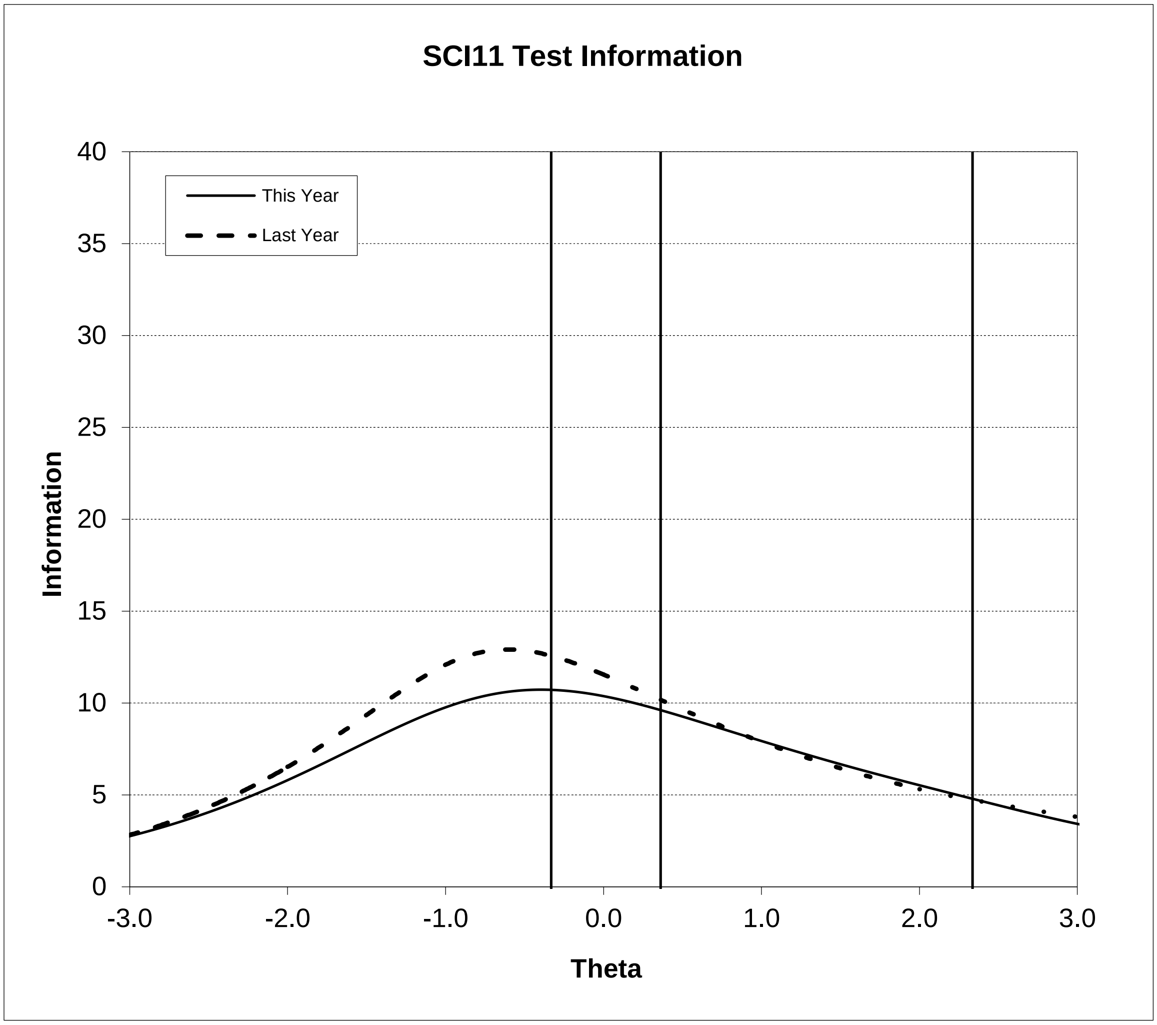


Figure I-2. 2013–14 MHSA: Test Information Function



APPENDIX J—RAW TO SCALED SCORE LOOK-UP TABLES

Table J-1. 2013–14 MHSA: Raw to Scaled Score Look-up Table—Science

Raw Score	*This year*			*Last year*		
	Scaled score	*Standard error*	*Performance level*	*Scaled score*	*Standard error*	*Performance level*
-13.33	1100	0	1	1100	0	1
-13	1100	0	1	1100	0	1
-12.67	1100	0	1	1100	0	1
-12.33	1100	0	1	1100	0	1
-12	1100	0	1	1100	0	1
-11.67	1100	0	1	1100	0	1
-11.33	1100	0	1	1100	0	1
-11	1100	0	1	1100	0	1
-10.67	1100	0	1	1100	0	1
-10.33	1100	0	1	1102	3	1
-10	1100	0	1	1102	3	1
-9.67	1102	3	1	1104	4	1
-9.33	1104	4	1	1106	4	1
-9	1104	4	1	1106	4	1
-8.67	1106	5	1	1108	5	1
-8.33	1106	5	1	1108	5	1
-8	1108	4	1	1110	4	1
-7.67	1108	4	1	1110	4	1
-7.33	1110	4	1	1112	3	1
-7	1110	4	1	1112	3	1
-6.67	1112	3	1	1112	3	1
-6.33	1112	3	1	1114	3	1
-6	1112	3	1	1114	3	1
-5.67	1114	3	1	1114	3	1
-5.33	1114	3	1	1114	3	1
-5	1114	3	1	1116	2	1
-4.67	1114	3	1	1116	2	1
-4.33	1116	2	1	1116	2	1
-4	1116	2	1	1118	2	1
-3.67	1116	2	1	1118	2	1
-3.33	1118	3	1	1118	2	1
-3	1118	3	1	1118	2	1
-2.67	1118	3	1	1118	2	1
-2.33	1118	3	1	1120	2	1
-2	1118	3	1	1120	2	1
-1.67	1120	2	1	1120	2	1
-1.33	1120	2	1	1120	2	1
-1	1120	2	1	1120	2	1
-0.67	1120	2	1	1122	2	1
-0.33	1122	2	1	1122	2	1
0	1122	2	1	1122	2	1
0.33	1122	2	1	1122	2	1
0.67	1122	2	1	1122	2	1
1	1122	2	1	1124	2	1
1.33	1122	2	1	1124	2	1

continued

Raw Score	This year			Last year		
	Scaled score	Standard error	Performance level	Scaled score	Standard error	Performance level
1.67	1124	2	1	1124	2	1
2	1124	2	1	1124	2	1
2.33	1124	2	1	1124	2	1
2.67	1124	2	1	1124	2	1
3	1124	2	1	1124	2	1
3.33	1124	2	1	1126	2	1
3.67	1126	2	1	1126	2	1
4	1126	2	1	1126	2	1
4.33	1126	2	1	1126	2	1
4.67	1126	2	1	1126	2	1
5	1126	2	1	1126	2	1
5.33	1126	2	1	1126	2	1
5.67	1128	2	1	1128	2	1
6	1128	2	1	1128	2	1
6.33	1128	2	1	1128	2	1
6.67	1128	2	1	1128	2	1
7	1128	2	1	1128	2	1
7.33	1128	2	1	1128	2	1
7.67	1128	2	1	1128	2	1
8	1130	2	1	1130	2	1
8.33	1130	2	1	1130	2	1
8.67	1130	2	1	1130	2	1
9	1130	2	1	1130	2	1
9.33	1130	2	1	1130	2	1
9.67	1130	2	1	1130	2	1
10	1130	2	1	1130	2	1
10.33	1130	2	1	1130	2	1
10.67	1132	1	1	1132	1	1
11	1132	1	1	1132	1	1
11.33	1132	1	1	1132	1	1
11.67	1132	1	1	1132	1	1
12	1132	1	1	1132	1	1
12.33	1132	1	1	1132	1	1
12.67	1132	1	1	1132	1	1
13	1132	1	1	1132	1	1
13.33	1132	1	1	1132	1	1
13.67	1132	1	1	1132	1	1
14	1132	1	1	1132	1	1
14.33	1132	1	1	1132	1	1
14.67	1132	1	1	1132	1	1
15	1132	1	1	1132	1	1
15.33	1132	1	1	1132	1	1
15.67	1136	3	2	1132	1	1
16	1136	3	2	1132	1	1
16.33	1136	3	2	1136	3	2
16.67	1136	3	2	1136	3	2
17	1136	3	2	1136	3	2
17.33	1136	3	2	1136	3	2

continued

Raw Score	*This year*			*Last year*		
	Scaled score	*Standard error*	*Performance level*	*Scaled score*	*Standard error*	*Performance level*
17.67	1136	3	2	1136	3	2
18	1136	3	2	1136	3	2
18.33	1138	2	2	1136	3	2
18.67	1138	2	2	1136	3	2
19	1138	2	2	1138	2	2
19.33	1138	2	2	1138	2	2
19.67	1138	2	2	1138	2	2
20	1138	2	2	1138	2	2
20.33	1138	2	2	1138	2	2
20.67	1138	2	2	1138	2	2
21	1140	2	2	1138	2	2
21.33	1140	2	2	1138	2	2
21.67	1140	2	2	1140	2	2
22	1140	2	2	1140	2	2
22.33	1140	2	2	1140	2	2
22.67	1140	2	2	1140	2	2
23	1140	2	2	1140	2	2
23.33	1140	2	2	1140	2	2
23.67	1140	2	2	1140	2	2
24	1140	2	2	1140	2	2
24.33	1140	2	2	1140	2	2
24.67	1142	2	3	1140	2	2
25	1142	2	3	1140	2	2
25.33	1142	2	3	1142	2	3
25.67	1142	2	3	1142	2	3
26	1144	2	3	1142	2	3
26.33	1144	2	3	1142	2	3
26.67	1144	2	3	1144	3	3
27	1144	2	3	1144	3	3
27.33	1144	2	3	1144	3	3
27.67	1144	2	3	1144	3	3
28	1144	2	3	1144	3	3
28.33	1146	2	3	1144	3	3
28.67	1146	2	3	1144	3	3
29	1146	2	3	1146	2	3
29.33	1146	2	3	1146	2	3
29.67	1146	2	3	1146	2	3
30	1146	2	3	1146	2	3
30.33	1146	2	3	1146	2	3
30.67	1148	2	3	1146	2	3
31	1148	2	3	1148	2	3
31.33	1148	2	3	1148	2	3
31.67	1148	2	3	1148	2	3
32	1148	2	3	1148	2	3
32.33	1148	2	3	1148	2	3
32.67	1150	2	3	1148	2	3
33	1150	2	3	1150	3	3
33.33	1150	2	3	1150	3	3

continued

Raw Score	*This year*			*Last year*		
	Scaled score	*Standard error*	*Performance level*	*Scaled score*	*Standard error*	*Performance level*
33.67	1150	2	3	1150	3	3
34	1150	2	3	1150	3	3
34.33	1150	2	3	1150	3	3
34.67	1150	2	3	1150	3	3
35	1152	2	3	1152	3	3
35.33	1152	2	3	1152	3	3
35.67	1152	2	3	1152	3	3
36	1152	2	3	1152	3	3
36.33	1152	2	3	1152	3	3
36.67	1154	2	3	1154	2	3
37	1154	2	3	1154	2	3
37.33	1154	2	3	1154	2	3
37.67	1154	2	3	1154	2	3
38	1154	2	3	1154	2	3
38.33	1154	2	3	1156	2	3
38.67	1156	2	3	1156	2	3
39	1156	2	3	1156	2	3
39.33	1156	2	3	1156	2	3
39.67	1156	2	3	1156	2	3
40	1156	2	3	1158	2	3
40.33	1158	2	3	1158	2	3
40.67	1158	2	3	1158	2	3
41	1158	2	3	1158	2	3
41.33	1158	2	3	1158	2	3
41.67	1160	3	3	1160	3	3
42	1160	3	3	1160	3	3
42.33	1160	3	3	1160	3	3
42.67	1160	3	3	1160	3	3
43	1160	3	3	1160	3	3
43.33	1160	3	3	1160	3	3
43.67	1160	3	3	1162	3	4
44	1162	3	4	1162	3	4
44.33	1164	4	4	1164	4	4
44.67	1164	4	4	1164	4	4
45	1164	4	4	1164	4	4
45.33	1164	4	4	1166	3	4
45.67	1166	3	4	1166	3	4
46	1166	3	4	1166	3	4
46.33	1166	3	4	1168	3	4
46.67	1168	3	4	1168	3	4
47	1168	3	4	1168	3	4
47.33	1168	3	4	1170	3	4
47.67	1170	4	4	1170	3	4
48	1170	4	4	1170	3	4
48.33	1170	4	4	1172	3	4
48.67	1172	4	4	1172	3	4
49	1172	4	4	1172	3	4
49.33	1174	4	4	1174	4	4

continued

Raw Score	This year			Last year		
	Scaled score	Standard error	Performance level	Scaled score	Standard error	Performance level
49.67	1174	4	4	1174	4	4
50	1176	4	4	1176	4	4
50.33	1176	4	4	1176	4	4
50.67	1178	3	4	1178	3	4
51	1178	3	4	1178	3	4
51.33	1180	0	4	1180	0	4
51.67	1180	0	4	1180	0	4
52	1180	0	4	1180	0	4
52.33	1180	0	4	1180	0	4
52.67	1180	0	4	1180	0	4
53	1180	0	4	1180	0	4
53.33	1180	0	4	1180	0	4
53.67	1180	0	4	1180	0	4
54	1180	0	4	1180	0	4
54.33	1180	0	4	1180	0	4
54.67	1180	0	4	1180	0	4
55	1180	0	4	1180	0	4
55.33	1180	0	4	1180	0	4
55.67	1180	0	4	1180	0	4
56	1180	0	4	1180	0	4

Table J-1. 2013–14 MHSA: Raw to Scaled Score Look-up Table—Mathematics

SAT Scaled Score	Scaled Score	Performance Level	Standard Error
200	1116	1	1
210	1118	1	1
220	1118	1	1
230	1120	1	2
240	1120	1	2
250	1122	1	2
260	1122	1	2
270	1124	1	3
280	1124	1	3
290	1126	1	3
300	1126	1	3
310	1128	1	3
320	1128	1	3
330	1130	1	3
340	1130	1	3
350	1132	1	3
360	1132	1	3
370	1134	2	4
380	1134	2	4
390	1136	2	4
400	1136	2	4
410	1138	2	4
420	1138	2	4
430	1140	2	4
440	1140	2	4
450	1140	2	4
460	1142	3	4
470	1144	3	4
480	1144	3	4
490	1146	3	4
500	1146	3	4
510	1148	3	4
520	1148	3	4
530	1150	3	4
540	1150	3	4
550	1152	3	4
560	1152	3	4
570	1154	3	4
580	1154	3	4
590	1156	3	4
600	1156	3	4
610	1158	3	4
620	1158	3	4
630	1160	3	4

continued

SAT Scaled Score	Scaled Score	Performance Level	Standard Error
640	1160	3	3
650	1160	3	3
660	1162	4	3
670	1164	4	3
680	1164	4	3
690	1166	4	3
700	1166	4	3
710	1168	4	3
720	1168	4	3
730	1170	4	3
740	1170	4	2
750	1172	4	2
760	1172	4	2
770	1174	4	2
780	1174	4	1
790	1176	4	1
800	1180	4	1

Table J-2. 2013–14 MHSA: Raw to Scaled Score Look-up Table—Reading

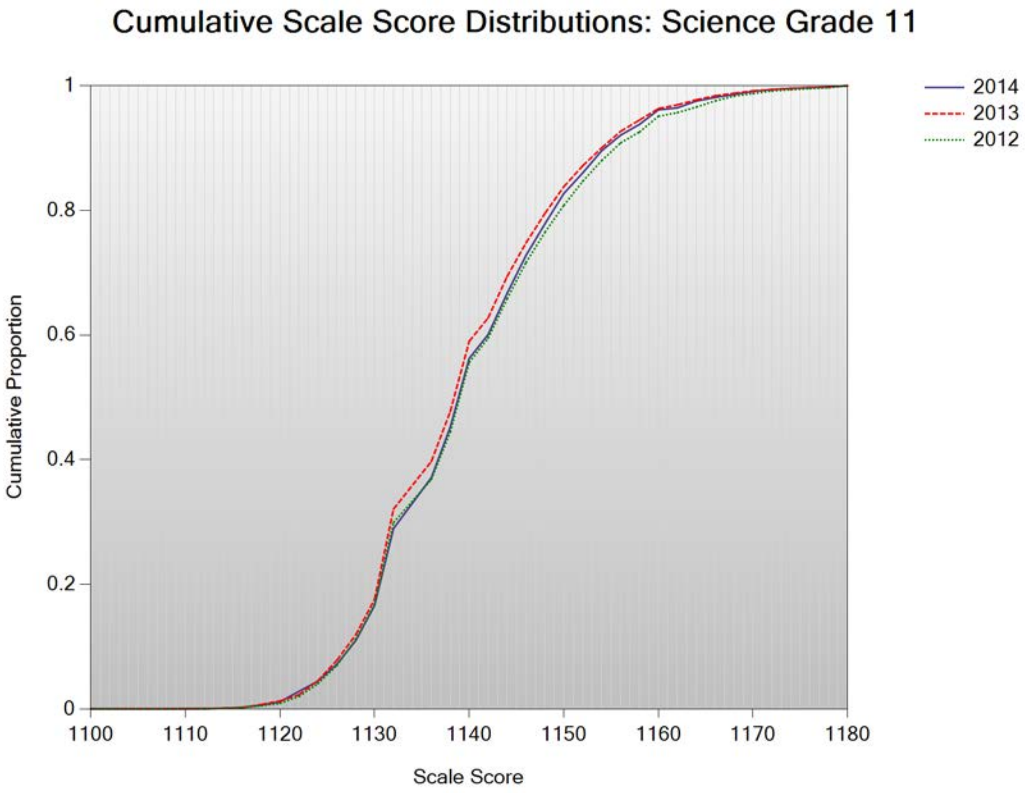
SAT Scaled Score	Scaled Score	Performance Level	Standard Error
200	1110	1	1
210	1110	1	1
220	1112	1	1
230	1114	1	3
240	1114	1	3
250	1116	1	3
260	1118	1	3
270	1118	1	4
280	1120	1	4
290	1120	1	4
300	1122	1	4
310	1124	1	4
320	1124	1	4
330	1126	1	4
340	1128	1	4
350	1128	1	4
360	1128	1	4
370	1130	2	5
380	1132	2	5
390	1134	2	5
400	1134	2	5
410	1136	2	5
420	1138	2	5
430	1138	2	5
440	1140	2	5
450	1140	2	5
460	1142	3	5
470	1144	3	5
480	1144	3	5
490	1146	3	5
500	1148	3	5
510	1148	3	5
520	1150	3	5
530	1150	3	5
540	1152	3	5
550	1154	3	5
560	1154	3	5
570	1156	3	5
580	1158	3	5
590	1158	3	5
600	1160	3	5
610	1160	3	5
620	1162	4	5
630	1164	4	5
640	1164	4	4
650	1166	4	4
660	1168	4	4
670	1168	4	4
680	1170	4	4
690	1170	4	4
700	1172	4	4
710	1174	4	4
720	1174	4	4
730	1176	4	4
740	1178	4	2
750	1178	4	2
760	1180	4	1
770	1180	4	1
780	1180	4	1
790	1180	4	1
800	1180	4	1

Table J-3. 2013–14 MHSA: Raw to Scaled Score Look-up Table—Writing

SAT Scaled Score	*Scaled Score*	*Performance Level*	*Standard Error*
200	1112	1	1
210	1114	1	1
220	1114	1	1
230	1116	1	2
240	1118	1	3
250	1118	1	3
260	1120	1	3
270	1120	1	4
280	1122	1	4
290	1124	1	4
300	1124	1	4
310	1126	1	4
320	1126	1	4
330	1128	1	4
340	1128	1	4
350	1130	2	4
360	1132	2	4
370	1132	2	5
380	1134	2	5
390	1134	2	5
400	1136	2	5
410	1138	2	5
420	1138	2	5
430	1140	2	5
440	1140	2	5
450	1142	3	5
460	1144	3	5
470	1144	3	5
480	1146	3	5
490	1146	3	5
500	1148	3	5
510	1150	3	5
520	1150	3	5
530	1152	3	5
540	1152	3	5
550	1154	3	5
560	1154	3	5
570	1156	3	5
580	1158	3	5
590	1158	3	5
600	1160	3	5
610	1160	3	5
620	1162	4	5
630	1164	4	5
640	1164	4	4
650	1166	4	4
660	1166	4	4
670	1168	4	4
680	1170	4	4
690	1170	4	4
700	1172	4	4
710	1172	4	4
720	1174	4	4
730	1174	4	4
740	1176	4	2
750	1178	4	2
760	1178	4	2
770	1180	4	1
780	1180	4	1
790	1180	4	1
800	1180	4	1

APPENDIX K—SCORE DISTRIBUTIONS

Figure K-1. 2013–14 MHSA: Score Distribution Plots—
Top: Science Bottom: Mathematics



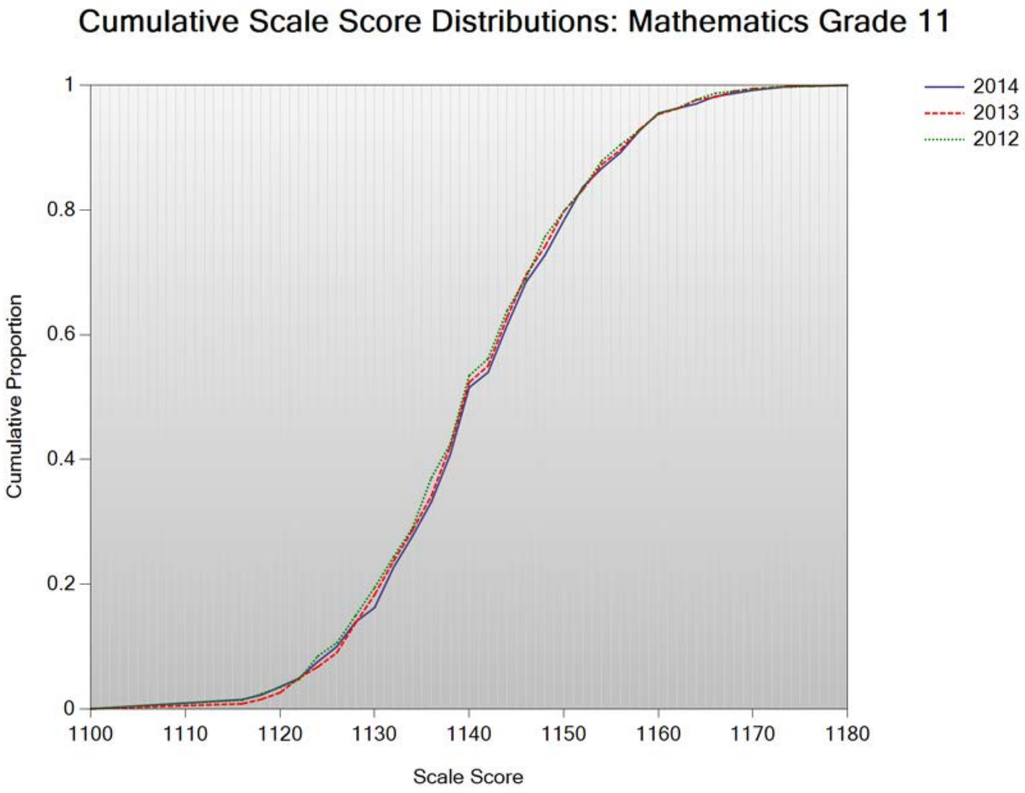
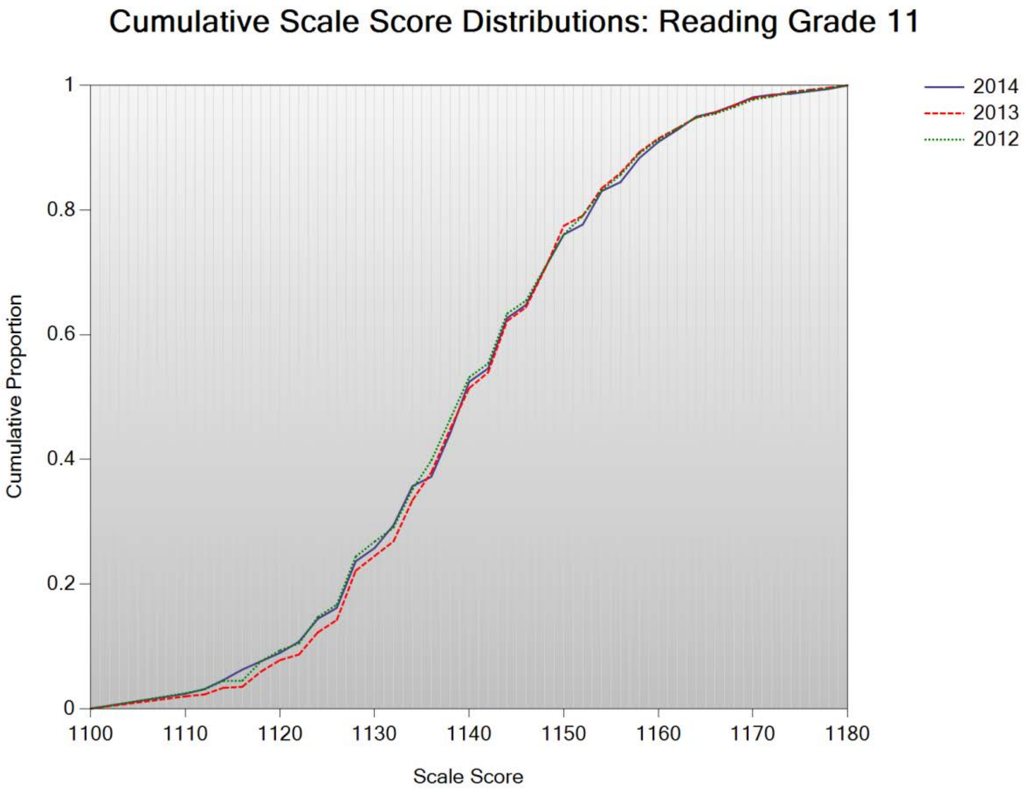


**Figure K-2. 2013–14 MHSA: Score Distribution Plot—
Top: Reading Bottom: Writing**



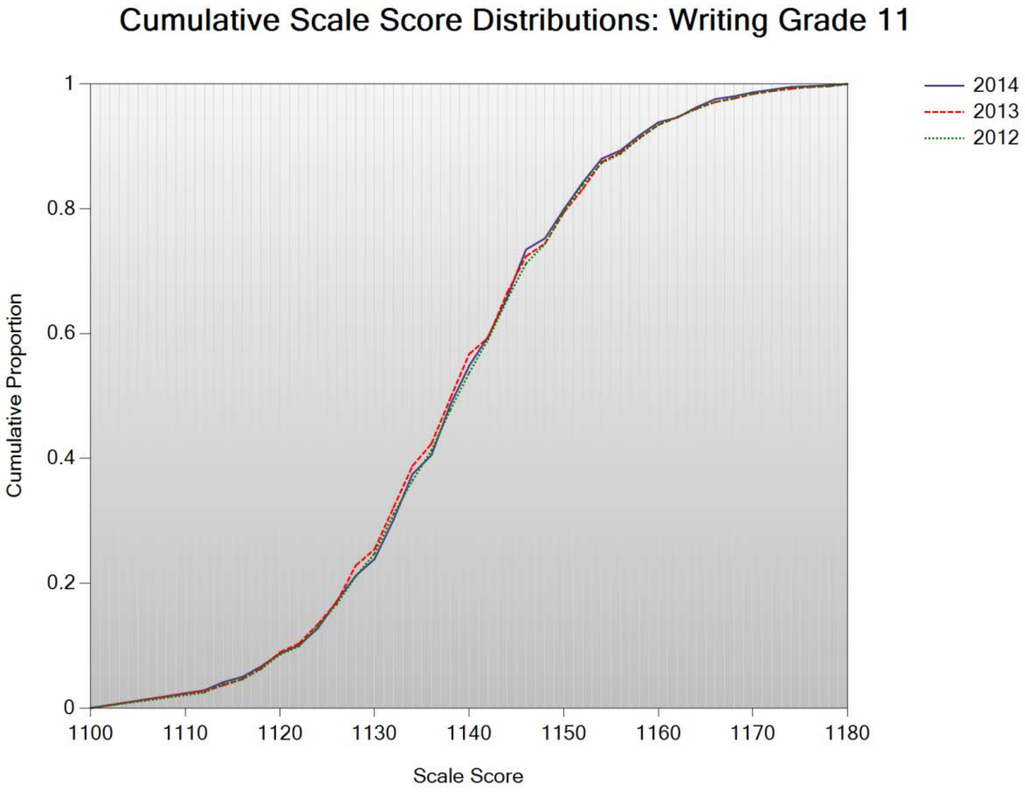


Table K-1. 2013–14 MHSA: Achievement Level Distributions—Science

Grade	*Achievement Level*	*Percent at Level*		
		2013–14	*2012–13*	*2011–12*
11	4	3.87	3.67	4.90
	3	39.89	37.35	39.50
	2	27.30	26.92	25.70
	1	28.94	32.06	29.90

Table K-2. 2013–14 MHSA: Achievement Level Distributions—Mathematics

Grade	*Achievement Level*	*Percent at Level*		
		2013–14	*2012–13*	*2011–12*
11	4	4.48	4.64	4.46
	3	43.98	43.00	42.09
	2	28.88	28.69	29.08
	1	22.65	23.67	24.36

Table K-3. 2013–14 MHSA: Achievement Level Distributions—Reading

Grade	*Achievement Level*	*Percent at Level*		
		2013–14	*2012–13*	*2011–12*
11	4	9.07	8.51	8.71
	3	38.49	40.07	38.12
	2	28.80	29.29	28.79
	1	23.63	22.13	24.38

Table K-4. 2013–14 MHSA: Achievement Level Distributions—Writing

Grade	*Achievement Level*	*Percent at Level*		
		2013–14	*2012–13*	*2011–12*
11	4	6.13	6.55	6.56
	3	39.03	36.72	39.73
	2	33.65	33.92	32.62
	1	21.19	22.81	21.09

APPENDIX L—RELIABILITY

Table L-1. 2013–14 MHSA: Subgroup Reliabilities—Science

Grade	*Group*	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	All Students	12,760	56	22.77	11.77	0.87	4.24
	Male	6,593	56	23.59	12.44	0.89	4.19
	Female	6,167	56	21.89	10.94	0.85	4.26
	Not Reported	0	56				
	Hispanic or Latino	185	56	20.86	11.53	0.87	4.22
	American Indian or Alaskan Native	92	56	19.05	10.05	0.83	4.20
	Asian	164	56	25.43	11.76	0.87	4.22
	Black or African American	405	56	14.46	10.92	0.85	4.22
	Native Hawaiian or Pacific Islander	14	56	23.14	11.59	0.85	4.52
	White (non-Hispanic)	11,774	56	23.08	11.70	0.87	4.24
	Two or more races	126	56	22.97	11.16	0.86	4.24
	Currently LEP students	247	56	9.35	8.04	0.74	4.12
	Former LEP student – monitoring year 1	35	56	16.17	7.88	0.71	4.21
	Former LEP student – monitoring year 2	59	56	19.01	7.79	0.70	4.26
	All Other Students	12,419	56	23.07	11.69	0.87	4.24
	Students with an IEP	1,678	56	12.25	9.89	0.83	4.12
	All Other Students	11,082	56	24.36	11.19	0.86	4.23
	Economically Disadvantaged Students	4,581	56	18.44	10.78	0.85	4.24
	All Other Students	8,179	56	25.20	11.60	0.87	4.22
	Migrant Students	2	56				
	All Other Students	12,758	56	22.77	11.77	0.87	4.24
	Students Receiving Title 1 Services	227	56	16.22	8.70	0.76	4.25
	All Other Students	12,533	56	22.89	11.78	0.87	4.24
	Students with a 504 plan	590	56	22.62	11.43	0.86	4.24
	All Other Students	12,170	56	22.78	11.78	0.87	4.24

Table L-2. 2013–14 MHSA: Reliabilities by Reporting Category—Science

Grade	*Item Reporting Category*	*Number of Items*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	Matter/Energy/Force/Motion	19	22	8.25	4.95	0.70	2.69
	Space/Earth	9	12	4.64	3.20	0.61	2.00
	The Living Environment	16	22	9.89	5.12	0.75	2.57
	The Physical Setting	28	34	12.88	7.41	0.79	3.36

Table L-3. 2013–14 MHSA: Subgroup Reliabilities—Mathematics

Grade	*Group*	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	All Students	11,837	54	21.82	13.14	0.93	3.36
	Male	6,006	54	22.57	13.72	0.94	3.36
	Female	5,831	54	21.05	12.48	0.93	3.36
	Not Reported	0	54				
	Hispanic or Latino	170	54	18.59	13.36	0.94	3.35
	American Indian or Alaskan Native	76	54	16.89	10.83	0.90	3.45
	Asian	168	54	27.14	14.64	0.95	3.25
	Black or African American	353	54	13.56	10.73	0.90	3.35
	Native Hawaiian or Pacific Islander	13	54	26.62	12.48	0.93	3.34
	White (non-Hispanic)	10,939	54	22.10	13.09	0.93	3.36
	Two or more races	118	54	20.40	12.47	0.93	3.39
	Currently LEP students	199	54	9.91	10.64	0.90	3.34
	Former LEP student – monitoring year 1	33	54	14.60	9.78	0.90	3.15
	Former LEP student – monitoring year 2	61	54	17.47	10.58	0.90	3.32
	All Other Students	11,544	54	22.07	13.10	0.93	3.36
	Students with an IEP	1,184	54	8.35	9.49	0.87	3.38
	All Other Students	10,653	54	23.32	12.63	0.93	3.35
	Economically Disadvantaged Students	4,069	54	16.41	11.36	0.91	3.42
	All Other Students	7,768	54	24.65	13.12	0.94	3.33
	Migrant Students	3	54				
	All Other Students	11,834	54	21.82	13.14	0.93	3.36
	Students Receiving Title 1 Services	214	54	13.01	8.55	0.84	3.45
	All Other Students	11,623	54	21.98	13.16	0.93	3.36
	Students with a 504 plan	540	54	20.36	12.00	0.92	3.40
	All Other Students	11,297	54	21.89	13.19	0.94	3.36

Table L-4. 2013–14 MHSA: Reliabilities by Reporting Category—Mathematics

Grade	*Item Reporting Category*	*Number of Items*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	Data, Statistics & Probability	7	7	2.86	2.12	0.65	1.26
	Functions & Algebra	19	19	7.05	4.98	0.85	1.94
	Geometry & Measurement	16	16	6.34	4.13	0.80	1.86
	Numbers & Operations	12	12	5.57	3.16	0.75	1.57

Table L-5. 2013–14 MHSA: Subgroup Reliabilities—Reading

Grade	*Group*	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	All Students	11,847	67	26.93	15.43	0.93	4.16
	Male	6,009	67	26.18	16.08	0.93	4.16
	Female	5,838	67	27.71	14.68	0.92	4.15
	Not Reported	0	67				
	Hispanic or Latino	170	67	24.50	15.46	0.93	4.19
	American Indian or Alaskan Native	76	67	22.36	13.58	0.90	4.24
	Asian	168	67	27.65	17.48	0.95	4.10
	Black or African American	353	67	18.67	14.36	0.92	4.15
	Native Hawaiian or Pacific Islander	13	67	30.98	15.55	0.93	4.11
	White (non-Hispanic)	10,949	67	27.26	15.36	0.93	4.16
	Two or more races	118	67	26.62	15.54	0.93	4.15
	Currently LEP students	199	67	8.71	9.34	0.81	4.05
	Former LEP student – monitoring year 1	33	67	15.40	8.44	0.78	3.99
	Former LEP student – monitoring year 2	61	67	20.32	9.07	0.78	4.21
	All Other Students	11,554	67	27.32	15.34	0.93	4.16
	Students with an IEP	1,190	67	12.22	12.90	0.89	4.19
	All Other Students	10,657	67	28.58	14.80	0.92	4.15
	Economically Disadvantaged Students	4,076	67	21.08	14.09	0.91	4.22
	All Other Students	7,771	67	30.00	15.21	0.93	4.12
	Migrant Students	3	67				
	All Other Students	11,844	67	26.94	15.42	0.93	4.16
	Students Receiving Title 1 Services	214	67	15.41	10.60	0.84	4.25
	All Other Students	11,633	67	27.15	15.42	0.93	4.16
	Students with a 504 plan	541	67	26.58	15.15	0.92	4.19
	All Other Students	11,306	67	26.95	15.44	0.93	4.16

Table L-6. 2013–14 MHSA: Reliabilities by Reporting Category—Reading

Grade	*Item Reporting Category*	*Number of Items*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	Informational	35	35	13.30	8.92	0.88	3.03
	Literary	9	9	4.43	2.69	0.67	1.54
	Word ID/Vocabulary	23	23	9.20	5.07	0.78	2.37

Table L-7. 2013–14 MHSA: Subgroup Reliabilities—Writing

Grade	*Group*	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	All Students	11,830	61	27.12	11.90	0.89	3.97
	Male	5,998	61	25.80	12.18	0.89	4.02
	Female	5,832	61	28.48	11.45	0.89	3.88
	Not Reported	0	61				
	Hispanic or Latino	168	61	24.43	12.24	0.89	4.09
	American Indian or Alaskan Native	78	61	22.24	10.94	0.86	4.03
	Asian	165	61	29.00	13.62	0.92	3.86
	Black or African American	343	61	20.95	10.92	0.87	3.93
	Native Hawaiian or Pacific Islander	13	61	27.96	12.50	0.89	4.14
	White (non-Hispanic)	10,945	61	27.38	11.85	0.89	3.97
	Two or more races	118	61	25.41	11.05	0.87	4.02
	Currently LEP students	180	61	14.18	7.50	0.73	3.89
	Former LEP student – monitoring year 1	33	61	18.96	7.09	0.73	3.69
	Former LEP student – monitoring year 2	61	61	22.34	7.93	0.75	3.95
	All Other Students	11,556	61	27.37	11.86	0.89	3.97
	Students with an IEP	1,191	61	14.57	9.47	0.81	4.14
	All Other Students	10,639	61	28.52	11.31	0.88	3.90
	Economically Disadvantaged Students	4,066	61	22.23	10.64	0.85	4.07
	All Other Students	7,764	61	29.67	11.72	0.89	3.88
	Migrant Students	3	61				
	All Other Students	11,827	61	27.12	11.90	0.89	3.97
	Students Receiving Title 1 Services	214	61	18.95	8.22	0.75	4.09
	All Other Students	11,616	61	27.27	11.91	0.89	3.97
	Students with a 504 plan	541	61	26.34	11.32	0.88	3.98
	All Other Students	11,289	61	27.15	11.93	0.89	3.97

Table L-8. 2013–14 MHSA: Reliabilities by Reporting Category—Writing

Grade	*Item Reporting Category*	*Number of Items*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
11	Revision in Context	6	6	1.99	2.01	0.64	1.21
	Sentence Correction	25	25	10.59	5.79	0.81	2.53
	Usage	18	18	8.13	4.38	0.75	2.18
	Writing Essay	1	12	6.41	1.82		

APPENDIX M—INTERRATER AGREEMENT SCIENCE

Table M-1. 2013–14 MHSA: Item-Level Interrater Consistency Statistics—Science

Grade	*Item Number*	*Number of*		*Percent*		*Correlation*	*Percent of Third Scores*
		Score Categories	*Responses Scored Twice*	*Exact*	*Adjacent*		
11	236887	5	1,258	52.23	42.37	0.72	4.93
	248437	5	1,120	57.05	36.07	0.78	5.80
	248439	5	1,271	67.11	24.70	0.75	8.65
	248504	5	1,225	65.63	30.69	0.76	3.35

APPENDIX N—SAMPLE REPORTS

Important Information for Parents/Guardians High School Assessment May 2014 Administration

Maine High School Assessment

STATE OF MAINE
DEPARTMENT OF EDUCATION
23 State House Station
Augusta, ME 04333
August 2014
James E. Rier, Jr.
COMMISSIONER OF EDUCATION

Dear Parents and Guardians,

The Maine High School Assessment is the State's measure of student progress in achieving the State standards known as *Learning Results*. It consists of the SAT Reasoning Test^{TM} (SAT) and a science test, and is administered to students in their third year of high school for state and federal accountability purposes (see graphic on the next page).

The Maine High School Assessment Report includes information on how your student scored on the SAT Reasoning^{TM} and Science tests administered in the spring of 2014, along with data on your child's school, school administrative unit, and state results. These results reflect scores based on MHSA questions that were taken by the nearly 14,000 students who were enrolled in their third year of high school across all Maine public schools. The MHSA employs an assessment design that requires students to create an essay response to a writing prompt, generate answers to open-ended mathematics and science questions, and select answers to multiple-choice questions in all four disciplines. More information about the design, history, and use of the SAT can be found at: http://www.maine.gov/doe/mhsa.

These results are reported across four achievement levels that describe the quality of your student's performance within each of the reading, mathematics, writing and science tests. These achievement level results are Maine-specific information not contained in any previously released SAT reports that your student may have received from the College Board. All scores contained in these reports are included for Maine reporting purposes only. While scores for most students may also be used for college admission, they may not be used for that purpose if a student received accommodations during the test administration that exceeded those made available by the College Board.

The Maine High School Assessment results should be viewed as one measure of student performance together with multiple local measures such as portfolios, performance exhibits, end-of-term grades, etc. to create a more complete picture of a student's overall academic performance. The staff at your school will be able to provide further information about your student's performance on the MHSA as well as your school's performance.

We hope you find this report informative as we continue to work toward improving the quality and effectiveness of instructional opportunities so that all Maine youth will graduate from high school prepared for college, career, and citizenship.

Sincerely,

James E. Rier, Jr.
Commissioner of Education

Information on the Maine High School Assessment

- More information about Maine's 2007 *Learning Results: Parameters for Essential Instruction* can be found at www.maine.gov/education/standards.htm.
- More information about the Maine SAT Initiative can be found at www.maine.gov/doe/mhsa.
- School reports, which allow you to review the Maine High School Assessment results by school, may be viewed at www.maine.gov/doe/mhsa as soon as they are available for posting.

Achievement Level Definitions

On this assessment, results are reported across four achievement levels. The general definitions below describe the quality of student work for each achievement level.

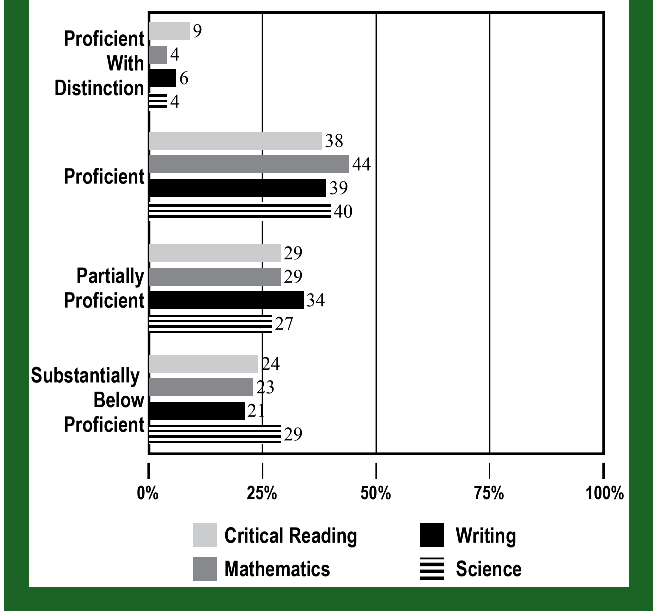
Proficient with Distinction: The student's work demonstrates in-depth understanding of essential concepts in a content area, including the ability to make multiple connections among central ideas. The student's responses demonstrate the ability to synthesize information, analyze and solve difficult problems, and apply complex concepts.

Proficient: The student's work demonstrates an understanding of essential concepts in a content area, including the ability to make connections among central ideas. The student's reponses demonstrate the ability to analyze and solve problems and apply concepts.

Partially Proficient: The student's work demonstrates incomplete understanding of essential concepts in a content area and inconsistent connections among central ideas. The student's responses demonstrate some ability to analyze and solve problems and apply concepts.

Substantially Below Proficient: The student's work demonstrates limited understanding of essential concepts in a content area and infrequent or inaccurate connections among central ideas. The student's responses demonstrate minimal ability to solve problems and apply concepts.

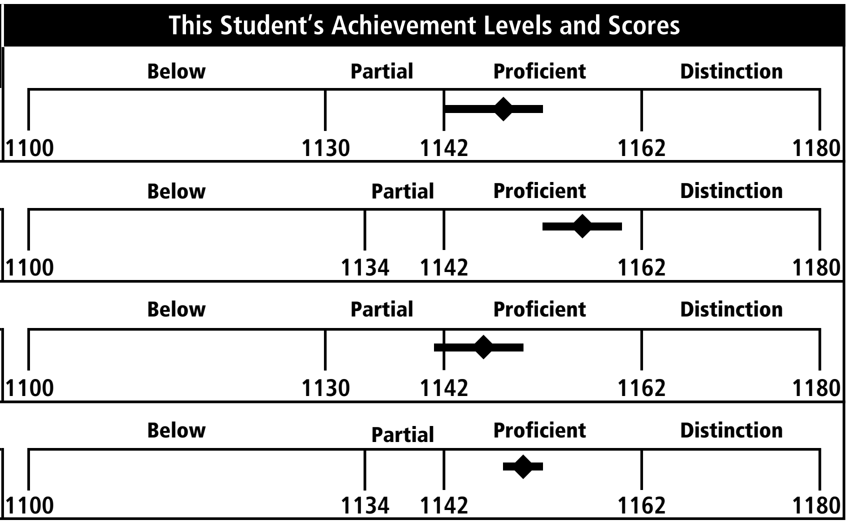
Maine High School Assessment Summary Results May 2014 Administration



Student	Grade	School	SAU
Kaitlyneliza X Majanosandoval	High School	Demonstration School 3	Demonstration District A

Content Area	Achievement Level	Score
Critical Reading	Proficient	1148
Mathematics	Proficient	1156
Writing	Proficient	1146
Science	Proficient	1150

This Student's Achievement Levels and Scores



See reverse side for description of achievement levels and state summary results.

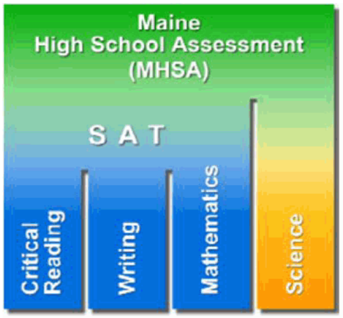
The diamond (♦) represents the student's score. The bar (▬) surrounding the score represents the probable range of scores for the student if he or she were to be tested many times. This range is based on a statistic called the standard error of measurement.

The scaled scores provided in the tables above reflect the 80-point scale used in all grades throughout the MeCAS system. The first two digits (11) denote the grade level of the assessment while the last two digits (00-80) show where the student scored on the 80-point scale. If your child took the SAT under an approved College Board administration, he or she should have received college reportable scores directly from the College Board approximately three weeks after testing. A conversion table showing all SAT scores in reading and writing and their MHSA equivalents can be found on the State's MHSA web page at http://www.maine.gov/doe/mhsa.

This Student's Achievement Level Relative to Student Achievement for School, SAU, and State

	Critical Reading				Mathematics				Writing				Science			
	Student	School	SAU	State	Student	School	SAU	State	Student	School	SAU	State	Student	School	SAU	State
Proficient with Distinction			11%	9%			6%	4%			4%	6%			5%	4%
Proficient	✓		39%	38%	✓		44%	44%	✓		46%	39%	✓		45%	40%
Partially Proficient			28%	29%			30%	29%			29%	34%			21%	27%
Substantially Below Proficient			23%	24%			20%	23%			21%	21%			29%	29%

Science	Total Possible Points	Student Points Earned	Average Points Earned		
			School	SAU	State
D. The Physical Setting Total	34	21.67		13.90	12.90
D1/D2 Space/Earth	12	9.67		5.10	4.60
D3/D4 Matter/Energy/ Force/Motion	22	12.00		8.70	8.20
E. The Living Environment Total	22	12.00		10.00	9.90



The table on the far left displays subscores for science. Similar information for all other subject areas has been provided in the student report mailed to you by the College Board. Formula scoring of all MHSA multiple-choice questions results in 1 point for a correct answer and a partial-point deduction for an incorrect answer. The graphic to the immediate left depicts the compostition of the entire MHSA program.

Maine High School Assessment

Name: Blair, Brandon S.
State ID: D11100016
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	First Year LEP	
Mathematics:	Partially Proficient	1138
Writing:	Special Consideration	
Science:	Substantially Below Proficient	1120

Maine High School Assessment

Name: Blaskovich, Kiley A.
State ID: D11100120
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Partially Proficient	1138
Mathematics:	Partially Proficient	1140
Writing:	Partially Proficient	1136
Science:	Substantially Below Proficient	1132

Maine High School Assessment

Name: Brooks, Colton R.
State ID: D11100105
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Proficient with Distinction	1162
Mathematics:	Proficient	1146
Writing:	Proficient	1154
Science:	Proficient	1142

Maine High School Assessment

Name: Brown, Shannon N.
State ID: D11100067
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Proficient	1150
Mathematics:	Proficient	1148
Writing:	Proficient	1154
Science:	Proficient	1148

Maine High School Assessment

Name: Curtsinger, Mark G.
State ID: D11100088
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Proficient	1148
Mathematics:	Proficient	1144
Writing:	Proficient	1146
Science:	Proficient	1148

Maine High School Assessment

Name: D'Agostino, Jordan M.
State ID: D11100079
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Partially Proficient	1140
Mathematics:	Proficient	1152
Writing:	Proficient	1144
Science:	Proficient	1142

Maine High School Assessment

Name: Dearmond, Kayla
State ID: D11100101
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Proficient	1154
Mathematics:	Proficient	1144
Writing:	Proficient	1146
Science:	Proficient	1150

Maine High School Assessment

Name: Degidio, Vito
State ID: D11100011
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Proficient	1158
Mathematics:	Proficient with Distinction	1166
Writing:	Proficient	1152
Science:	Proficient	1152

Maine High School Assessment

Name: Doyle, Cameron B.
State ID: D11100118
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Substantially Below Proficient	1120
Mathematics:	Substantially Below Proficient	1124
Writing:	Substantially Below Proficient	1128
Science:	Substantially Below Proficient	1132

Maine High School Assessment

Name: Ehrlich, Daniel D.
State ID: D11100019
School: Demonstration School 1
SAU: Demonstration District A

Date: 05/2014

------------	Achievement Levels	---- Scaled Scores
Reading:	Substantially Below Proficient	1112
Mathematics:	Proficient	1146
Writing:	Substantially Below Proficient	1128
Science:	Partially Proficient	1140

August 2014

Maine Department of Education

2013-2014 School Year Reports

Dear School Board Members and School Personnel:

The Maine High School Assessment is the State's measure of student progress in achieving the State standards known as *Learning Results*. It consists of the SAT Reasoning Test™ (SAT) and a science test, and is administered to students in their third year of high school for state and federal purposes.

These Maine High School Assessment Summary Reports contain the results of your students' performance in critical reading, mathematics, writing, and science reported according to the academic standards described above and disaggregated by student and school characteristics. The MHSA achievement level standards for the critical reading, writing, mathematics, and science sections of the MHSA were determined by Maine educators with specific expertise within the content areas. This report, together with individual student and subject-specific student item-level reports, provides support for use in program evaluation and planning. All scores contained in these reports are included for Maine state and federal reporting purposes only. While scores from the SAT may also be used for college admission by most students, they may not be used for that purpose if a student received accommodations during the test administration that exceeded those made available by the College Board.

These results reflect scores based on SAT and science test questions that were taken by the nearly 14,000 publicly-funded students who were enrolled in their third year of high school across all Maine schools. The MHSA employs an assessment design that requires students to create an essay response to a writing prompt, generate answers to open-ended mathematics and science questions, and select answers to multiple-choice questions in all four disciplines. More information about the design, history, and use of the SAT can be found at: http://www.maine.gov/education/mhsa/index.htm.

I look forward to working with you in support of our continued efforts to improve the quality and effectiveness of the instructional opportunities designed to help all students achieve the high standards of the *Learning Results* and graduate from any Maine high school prepared for college, career, and citizenship.

Sincerely,

James E. Rier, Jr.
Commissioner of Education

Maine High School Assessment

SAU Report

Test Date: May 2014

Code: DEMA

SAU: Demonstration District A

Contents of the Report

The report is divided into seven main sections including a section describing the students tested and a separate section for the results in each content area.

Topic	*Page*
Summary of Scores	2
Summary of Student Participation	3
Critical Reading Results	4-5
Mathematics Results	6-7
Writing Results	8-9
Science Results	10-11
Questionnaire Science Results	12

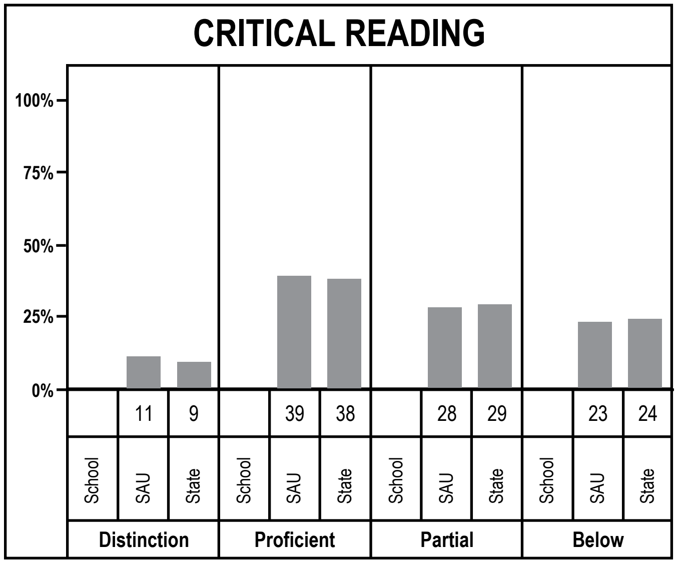
Maine High School Assessment

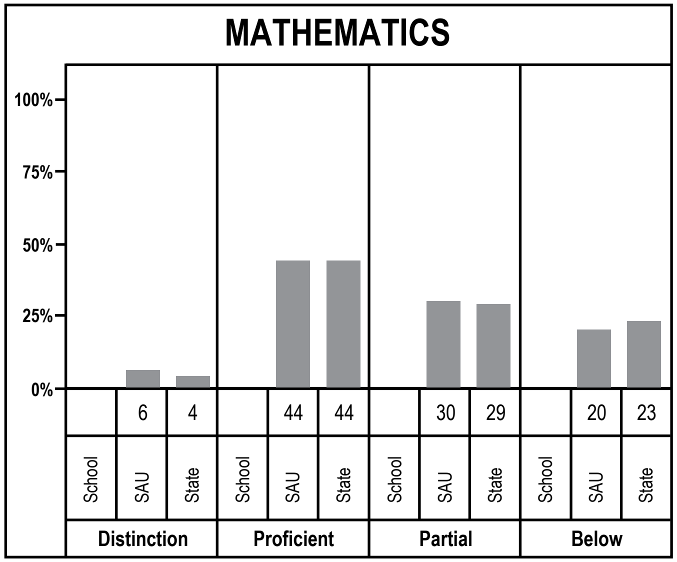
SUMMARY OF SCORES

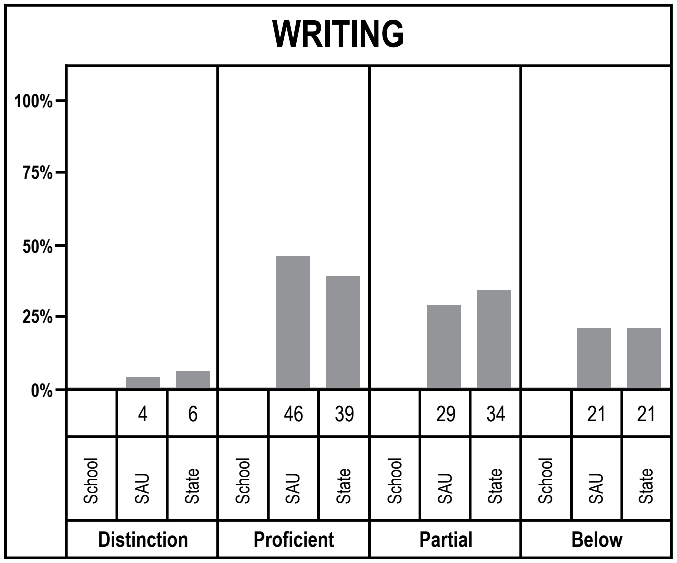
Test Date: May 2014
SAU: Demonstration District A

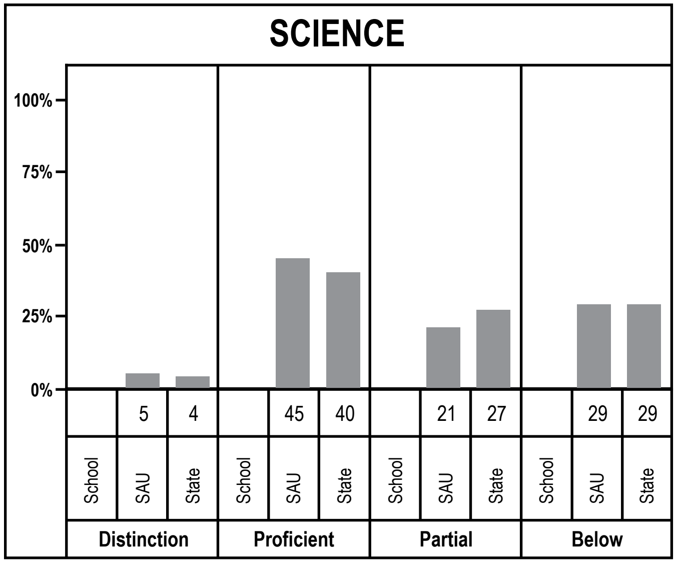
Summary of SAU and State Scores

Year	Average Scaled Score		
	School	SAU	State
Critical Reading			
2011–2012		1140	1141
2012–2013		1139	1141
2013–2014		**1142**	**1141**
Cum. Average*		1140	1141
Mathematics			
2011–2012		1139	1141
2012–2013		1140	1142
2013–2014		**1143**	**1142**
Cum. Average*		1141	1142
Writing			
2011–2012		1139	1140
2012–2013		1137	1140
2013–2014		**1141**	**1140**
Cum. Average*		1139	1140
Science			
2011–2012		1140	1141
2012–2013		1138	1140
2013–2014		**1142**	**1141**
Cum. Average*		1140	1141


CRITICAL READING


MATHEMATICS


WRITING


SCIENCE

*Cumulative averages are weighted–i.e., the scaled scores are averaged proportionally based on the numbers of students in each year.

Maine
High School
Assessment

SUMMARY OF STUDENT PARTICIPATION

Test Date: May 2014
SAU: Demonstration District A

CATEGORY OF PARTICIPATION		Enrollment[1] during testing window						CONTENT AREA PARTICIPATION[2]																							
								Critical Reading						Mathematics						Writing						Science					
		School		SAU		State		School		SAU		State		School		SAU		State		School		SAU		State		School		SAU		State	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Total number of students				127	100	13574	100			121	96	13031	96			121	96	13039	96			120	96	13009	96			121	96	12952	95
Ethnicity	Hispanic or Latino			1	1	192	1			1	100	187	97			1	100	189	98			1	100	184	97			1	100	188	98
Not Hispanic or Latino	American Indian or Alaskan Native			1	1	103	1			1	100	93	90			1	100	93	90			1	100	95	92			1	100	95	93
	Asian			4	3	178	1			4	100	176	99			4	100	176	99			4	100	172	99			4	100	167	94
	Black or African American			5	4	442	3			5	100	423	96			5	100	426	97			4	100	401	96			5	100	416	94
	Native Hawaiian or Pacific Islander			1	1	14	<1			1	100	14	100			1	100	14	100			1	100	14	100			1	100	14	100
	White			114	90	12512	92			108	96	12011	96			108	96	12014	96			108	96	12017	96			108	96	11945	96
	Two or more races			1	1	133	1			1	100	127	96			1	100	127	96			1	100	126	96			1	100	127	95
Identified disability				16	13	2051	15			15	94	1852	91			15	94	1853	91			15	94	1859	91			15	100	1870	91
Current LEP				2	2	285	2			2	100	270	95			2	100	271	95			1	100	232	94			2	100	258	91
Economically disadvantaged				52	41	4999	37			48	94	4688	94			48	94	4699	94			47	94	4675	94			48	94	4683	94
Migrant				1	1	4	<1			1	100	4	100			1	100	4	100			1	100	4	100			0	0	2	50

MODE OF PARTICIPATION[3]	Critical Reading						Mathematics						Writing						Science					
	School		SAU		State		School		SAU		State		School		SAU		State		School		SAU		State	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Participation without accommodations			108	85	11522	85			108	85	11514	85			108	85	11504	85			109	86	11512	85
Identified disability (IEP)			6	6	765	7			6	6	761	7			6	6	765	7			7	6	821	7
LEP			1	1	188	2			1	1	188	2			1	1	169	1			1	1	183	2
Participation with accommodations			11	9	1299	10			12	9	1331	10			11	9	1313	10			11	9	1249	9
Identified disability (IEP)			8	73	892	69			8	67	898	67			8	73	902	69			7	64	858	69
LEP			0	0	56	4			1	8	72	5			0	0	52	4			1	9	64	5
Participation through alternate assessment (PAAP)			1	1	195	1			1	1	194	1			1	1	192	1			1	1	191	1
Identified disability (IEP)			1	100	195	100			1	100	194	100			1	100	192	100			1	100	191	100
LEP			0	0	11	6			0	0	11	6			0	0	11	6			0	0	11	6
Approved non-participation in reading – 1st year LEP			1	1	15	<1																		
Approved non-participation – special consideration			1	1	20	<1			1	1	20	<1			2	2	58	<1			1	1	10	<1
Non-participation – other			5	4	523	4			5	4	515	4			5	4	507	4			5	4	612	5

[1] Percents are the percentage of students enrolled in each participation category.

[2] Percents are the percentage of students, including those who participated through alternate assessment (PAAP), who participated in the content area.

[3] Percents are the percentage of students in each content area by mode.

Maine
High School
Assessment

CRITICAL READING RESULTS

Test Date: May 2014
SAU: Demonstration District A

ACHIEVEMENT LEVELS: Achievement level definitions describe the quality of a student's responses on state-level assessments in relation to the reading standards for achieving Maine's *Learning Results*.

Maine state-level assessments measure the knowledge and skills of students by sampling identified standards within reading at the grade level assessed. Evidence includes responses to multiple-choice items in an "on demand" setting.

		STUDENTS AT EACH ACHIEVEMENT LEVEL					
		School		SAU		State	
		N	%	N	%	N	%
Proficient with Distinction – The student's work demonstrates the ability to read and interpret literary and informational texts appropriate for the grade level by applying a variety of reasoning skills and prior knowledge as the student draws in-depth inferences, analyzes texts for subtle clues, synthesizes information across texts, and uses knowledge of text structures and literary devices to make deeper connections within or across texts to increase comprehension. (Scaled Score 1162-1180)	2011–2012			8	7	1,156	9
	2012–2013			8	7	1,096	9
	2013–2014			**13**	**11**	**1,163**	**9**
	Cum. Average*			29	8	3,415	9
Proficient– The student's work demonstrates the ability to read and interpret literary and informational texts appropriate for the grade level by applying a variety of reasoning skills and prior knowledge as the student draws inferences, identifies summary statements, connects ideas within and across texts, and uses knowledge of text structures and literary devices to increase comprehension. (Scaled Score 1142-1160)	2011–2012			45	38	5,057	38
	2012–2013			37	33	5,159	40
	2013–2014			**46**	**39**	**4,935**	**38**
	Cum. Average*			128	37	15,151	39
Partially Proficient – The student's work demonstrates an inconsistent ability to read and interpret literary and informational texts appropriate for the grade level. The student's ability to use a variety of reasoning skills and prior knowledge varies depending on the texts as s/he draws inferences, identifies summary statements, connects ideas within and across texts, and uses knowledge of text structures and literary devices to support comprehension. (Scaled Score 1130-1140)	2011–2012			37	31	3,820	29
	2012–2013			29	26	3,768	29
	2013–2014			**33**	**28**	**3,693**	**29**
	Cum. Average*			99	28	11,281	29
Substantially Below Proficient – The student's work demonstrates a limited ability to read and interpret literary and informational texts appropriate for the grade level. The student's responses are often incorrect leaving the impression that the student found it difficult to use a variety of reasoning skills and prior knowledge as s/he draws inferences, identifies summary statements, connects ideas within and across texts, or uses knowledge of text structures and literary devices to support comprehension. (Scaled Score 1100-1128)	2011–2012			30	25	3,234	24
	2012–2013			37	33	2,840	22
	2013–2014			**27**	**23**	**3,030**	**24**
	Cum. Average*			94	27	9,104	23

* Percentages are calculated by dividing the cumulative total of the number of students in the achievement level by the cumulative total of the number of students tested.

Maine
High School
Assessment

CRITICAL READING RESULTS BY REPORTING SUBGROUPS

Test Date: May 2014
SAU: Demonstration District A

REPORTING CATEGORIES	SAU													State											
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
All Students	**127**	**3**	**5**	**119**	**13**	**11**	**46**	**39**	**33**	**28**	**27**	**23**	**1142**	**12,821**	**9**	**38**	**29**	**24**	**1141**						
Gender																									
Male	64	3	2	59	7	12	24	41	15	25	13	22	1142	6,592	10	35	28	27	1140						
Female	63	0	3	60	6	10	22	37	18	30	14	23	1142	6,229	8	42	30	20	1142						
Not Reported	0	0	0	0										0											
Primary Race/Ethnicity																									
Hispanic or Latino	1	0	0	1										183	7	37	30	26	1139						
Not Hispanic or Latino																									
American Indian or Alaskan Native	1	0	0	1										90	2	30	34	33	1136						
Asian	4	0	0	4										173	14	32	29	24	1142						
Black or African American	5	1	0	4										403	2	23	28	47	1133						
Native Hawaiian or Pacific Islander	1	0	0	1										14	14	43	29	14	1144						
White (non-Hispanic)	114	2	5	107	12	11	38	36	31	29	26	24	1142	11,832	9	39	29	23	1141						
Two or more races	1	0	0	1										126	10	39	25	26	1141						
LEP Status																									
Currently LEP student	2	1	0	1										244	0	5	17	78	1123						
Former LEP student - monitoring year 1	1	0	0	1										35	0	6	54	40	1132						
Former LEP student - monitoring year 2	1	0	0	1										61	0	30	49	21	1136						
All Other Students	123	2	5	116	13	11	46	40	31	27	26	22	1142	12,481	9	39	29	23	1141						
IEP																									
Students with an IEP	16	1	1	14	0	0	2	14	2	14	10	71	1127	1,657	2	11	21	67	1127						
All Other Students	111	2	4	105	13	12	44	42	31	30	17	16	1144	11,164	10	43	30	17	1143						
SES																									
Economically Disadvantaged Students	52	2	3	47	2	4	16	34	13	28	16	34	1137	4,574	3	28	32	36	1135						
All Other Students	75	1	2	72	11	15	30	42	20	28	11	15	1145	8,247	12	44	27	17	1144						
Migrant																									
Migrant Students	1	0	0	1										4											
All Other Students	126	3	5	118	13	11	46	39	32	27	27	23	1142	12,817	9	39	29	24	1141						
Title 1																									
Students Receiving Title 1 Services	1	0	0	1										231	<1	15	39	46	1131						
All Other Students	126	3	5	118	13	11	46	39	32	27	27	23	1142	12,590	9	39	29	23	1141						
504 Plan																									
Students with a 504 plan	5	0	0	5										598	10	36	31	23	1141						
All Other Students	122	3	5	114	13	11	44	39	32	28	25	22	1142	12,223	9	39	29	24	1141						

Level 4 = Proficient with Distinction Level 3 = Proficient Level 2 = Partially Proficient Level 1 = Substantially Below Proficient

Note: Some achievement level results have been left blank because fewer than ten (10) students were tested. **N** = Number

Maine
High School
Assessment

MATHEMATICS RESULTS

Test Date: May 2014
SAU: Demonstration District A

ACHIEVEMENT LEVELS: Achievement level definitions describe the quality of a student’s responses on state-level assessments in relation to the mathematics standards for achieving Maine’s *Learning Results*.

Maine state-level assessments measure the knowledge and skills of students by sampling identified standards within mathematics at the grade level assessed. Evidence includes responses to a combination of multiple-choice items and items requiring student-created responses in an “on demand” setting.

		STUDENTS AT EACH ACHIEVEMENT LEVEL					
		School		SAU		State	
		N	%	N	%	N	%
Proficient with Distinction – The student’s work demonstrates in-depth understanding of essential concepts in mathematics, including the ability to make multiple connections among central ideas. The student’s responses demonstrate the ability to synthesize information, analyze and solve difficult or unfamiliar problems, and apply complex concepts. (Scaled Score 1162–1180)	2011–2012			2	2	592	4
	2012–2013			4	4	599	5
	2013–2014			**7**	**6**	**576**	**4**
	Cum. Average*			13	4	1,767	5
Proficient– The student’s work demonstrates an understanding of essential concepts in mathematics, including the ability to make connections among central ideas. The student’s responses demonstrate the ability to reason, analyze and solve problems, and apply concepts. (Scaled Score 1142–1160)	2011–2012			46	38	5,586	42
	2012–2013			42	38	5,544	43
	2013–2014			**53**	**44**	**5,649**	**44**
	Cum. Average*			141	40	16,779	43
Partially Proficient – The student’s work demonstrates incomplete understanding of essential concepts in mathematics and inconsistent connections among central ideas. The student’s responses demonstrate some ability to analyze and solve problems and apply concepts. (Scaled Score 1134–1140)	2011–2012			40	33	3,859	29
	2012–2013			34	31	3,692	29
	2013–2014			**36**	**30**	**3,710**	**29**
	Cum. Average*			110	31	11,261	29
Substantially Below Proficient – The student’s work demonstrates limited understanding of essential concepts in mathematics and infrequent or inaccurate connections among central ideas. The student’s responses demonstrate minimal ability to solve problems and apply concepts. (Scaled Score 1100–1132)	2011–2012			32	27	3,233	24
	2012–2013			31	28	3,037	24
	2013–2014			**24**	**20**	**2,910**	**23**
	Cum. Average*			87	25	9,180	24

* Percentages are calculated by dividing the cumulative total of the number of students in the achievement level by the cumulative total of the number of students tested.

Maine
High School
Assessment

MATHEMATICS RESULTS
BY REPORTING SUBGROUPS

Test Date: May 2014
SAU: Demonstration District A

REPORTING CATEGORIES	SAU													State											
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
All Students	**127**	**2**	**5**	**120**	**7**	**6**	**53**	**44**	**36**	**30**	**24**	**20**	**1143**	**12,845**	**4**	**44**	**29**	**23**	**1142**						
Gender																									
Male	64	2	2	60	4	7	26	43	20	33	10	17	1144	6,609	6	44	27	23	1142						
Female	63	0	3	60	3	5	27	45	16	27	14	23	1142	6,236	3	44	30	22	1141						
Not Reported	0	0	0	0										0											
Primary Race/Ethnicity																									
Hispanic or Latino	1	0	0	1										186	2	39	26	33	1139						
Not Hispanic or Latino																									
American Indian or Alaskan Native	1	0	0	1										90	0	29	37	34	1137						
Asian	4	0	0	4										174	14	50	21	15	1147						
Black or African American	5	0	0	5										415	<1	21	31	47	1134						
Native Hawaiian or Pacific Islander	1	0	0	1										14	7	57	21	14	1145						
White (non-Hispanic)	114	2	5	107	6	6	47	44	34	32	20	19	1143	11,840	5	45	29	22	1142						
Two or more races	1	0	0	1										126	4	38	33	25	1141						
LEP Status																									
Currently LEP student	2	0	0	2										260	1	12	23	64	1131						
Former LEP student - monitoring year 1	1	0	0	1										35	3	17	51	29	1137						
Former LEP student - monitoring year 2	1	0	0	1										61	0	34	43	23	1139						
All Other Students	123	2	5	116	6	5	53	46	35	30	22	19	1143	12,489	5	45	29	22	1142						
IEP																									
Students with an IEP	16	1	1	14	0	0	0	0	2	14	12	86	1128	1,659	<1	11	20	69	1130						
All Other Students	111	1	4	106	7	7	53	50	34	32	12	11	1145	11,186	5	49	30	16	1144						
SES																									
Economically Disadvantaged Students	52	1	3	48	1	2	19	40	11	23	17	35	1139	4,595	1	31	33	35	1137						
All Other Students	75	1	2	72	6	8	34	47	25	35	7	10	1146	8,250	7	51	26	16	1145						
Migrant																									
Migrant Students	1	0	0	1										4											
All Other Students	126	2	5	119	7	6	53	45	35	29	24	20	1143	12,841	4	44	29	23	1142						
Title 1																									
Students Receiving Title 1 Services	1	0	0	1										231	<1	17	43	39	1135						
All Other Students	126	2	5	119	7	6	53	45	35	29	24	20	1143	12,614	5	44	29	22	1142						
504 Plan																									
Students with a 504 plan	5	0	0	5										598	3	41	33	23	1141						
All Other Students	122	2	5	115	7	6	51	44	34	30	23	20	1143	12,247	5	44	29	23	1142						

Level 4 = Proficient with Distinction Level 3 = Proficient Level 2 = Partially Proficient Level 1 = Substantially Below Proficient
Note: Some achievement level results have been left blank because fewer than ten (10) students were tested. **N** = Number

Maine
High School
Assessment

WRITING RESULTS

Test Date: May 2014
SAU: Demonstration District A

ACHIEVEMENT LEVELS: Achievement level definitions describe the quality of a student's responses on state-level assessments in relation to the writing standards for achieving Maine's *Learning Results*.

Maine state-level assessments measure the knowledge and skills of students by sampling identified standards within writing at the grade level assessed. Evidence includes responses to a combination of multiple-choice items and items requiring student-created responses in an "on demand" setting.

		STUDENTS AT EACH ACHIEVEMENT LEVEL					
		School		SAU		State	
		N	%	N	%	N	%
Proficient with Distinction – The student's responses demonstrate skillful ability to select clear, precise sentence improvements that are free of awkwardness or ambiguity; to recognize grammar and usage errors; and to select revisions that add to the clarity, precision, and overall effectiveness of a passage. The student's essay demonstrates an effectively developed and insightful point of view on the issue and outstanding critical thinking, with clearly appropriate examples, reasons, and other evidence to support a position. The essay is well-organized and clearly focused, demonstrating clear coherence and smooth progression of ideas and free of most errors in grammar, usage, and mechanics. (Scaled Score 1162–1180)	2011–2012			8	7	871	7
	2012–2013			6	5	845	7
	2013–2014			**5**	**4**	**786**	**6**
	Cum. Average*			19	5	2,502	6
Proficient– The student's responses demonstrate ability to select clear sentence improvements that are free of awkwardness or ambiguity; to recognize grammar and usage errors; and to select revisions that add to the clarity and overall effectiveness of a passage. The student's essay demonstrates an effectively developed point of view on the issue and strong critical thinking, with generally appropriate examples, reasons, and other evidence to support a position. The essay is well-organized and focused, demonstrating coherence and progression of ideas and generally free of most errors in grammar, usage, and mechanics. (Scaled Score 1142–1160)	2011–2012			43	36	5,274	40
	2012–2013			34	30	4,733	37
	2013–2014			**55**	**46**	**5,002**	**39**
	Cum. Average*			132	38	15,009	39
Partially Proficient – The student's responses demonstrate inconsistent ability to select clear sentence improvements that are free of awkwardness or ambiguity; to recognize grammar and usage errors; and to select revisions that add to the clarity and overall effectiveness of a passage. The student's essay demonstrates a developed point of view on the issue and some critical thinking, but may do so inconsistently or with inadequate examples, reasons, or other evidence to support a position. The essay is generally organized and focused, but may demonstrate some lapses in coherence or progression of ideas and may contain errors in grammar, usage, and mechanics. (Scaled Score 1130–1140)	2011–2012			44	37	4,330	33
	2012–2013			36	32	4,369	34
	2013–2014			**34**	**29**	**4,313**	**34**
	Cum. Average*			114	32	13,012	33
Substantially Below Proficient – The student's responses demonstrate limited ability to select clear sentence improvements that are free of awkwardness or ambiguity; to recognize grammar and usage errors; and to select revisions that add to the clarity and overall effectiveness of a passage. The student's essay demonstrates a vague or seriously limited point of view on the issues and weak critical thinking, with inappropriate or insufficient examples, reasons, or other evidence to support a position. The essay is poorly organized and/or focused and may contain an accumulation of errors in grammar, usage, and mechanics that interfere with understanding the message of the essay. (Scaled Score 1100–1128)	2011–2012			25	21	2,800	21
	2012–2013			36	32	2,926	23
	2013–2014			**25**	**21**	**2,716**	**21**
	Cum. Average*			86	25	8,442	22

* Percentages are calculated by dividing the cumulative total of the number of students in the achievement level by the cumulative total of the number of students tested.

Maine
High School
Assessment

WRITING RESULTS
BY REPORTING SUBGROUPS

Test Date: May 2014
SAU: Demonstration District A

REPORTING CATEGORIES	SAU													State											
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
All Students	**127**	**3**	**5**	**119**	**5**	**4**	**55**	**46**	**34**	**29**	**25**	**21**	**1141**	**12,817**	**6**	**39**	**34**	**21**	**1140**						
Gender																									
Male	64	3	2	59	3	5	23	39	19	32	14	24	1141	6,593	5	34	34	26	1138						
Female	63	0	3	60	2	3	32	53	15	25	11	18	1141	6,224	7	44	33	16	1142						
Not Reported	0	0	0	0										0											
Primary Race/Ethnicity																									
Hispanic or Latino	1	0	0	1										181	4	31	36	28	1137						
Not Hispanic or Latino																									
American Indian or Alaskan Native	1	0	0	1										92	3	16	46	35	1134						
Asian	4	0	0	4										170	14	37	29	20	1143						
Black or African American	5	1	0	4										390	2	20	41	38	1133						
Native Hawaiian or Pacific Islander	1	0	0	1										14	7	36	29	29	1140						
White (non-Hispanic)	114	2	5	107	5	5	49	46	30	28	23	21	1141	11,845	6	40	33	20	1140						
Two or more races	1	0	0	1										125	3	32	38	27	1138						
LEP Status																									
Currently LEP student	2	1	0	1										221	0	3	34	63	1126						
Former LEP student - monitoring year 1	1	0	0	1										35	0	11	60	29	1133						
Former LEP student - monitoring year 2	1	0	0	1										61	0	25	54	21	1136						
All Other Students	123	2	5	116	5	4	55	47	33	28	23	20	1141	12,500	6	40	33	20	1140						
IEP																									
Students with an IEP	16	1	1	14	0	0	2	14	1	7	11	79	1128	1,667	<1	8	24	67	1126						
All Other Students	111	2	4	105	5	5	53	50	33	31	14	13	1143	11,150	7	44	35	14	1142						
SES																									
Economically Disadvantaged Students	52	2	3	47	0	0	18	38	14	30	15	32	1137	4,572	1	27	39	33	1134						
All Other Students	75	1	2	72	5	7	37	51	20	28	10	14	1143	8,245	9	46	31	14	1143						
Migrant																									
Migrant Students	1	0	0	1										4											
All Other Students	126	3	5	118	5	4	54	46	34	29	25	21	1141	12,813	6	39	34	21	1140						
Title 1																									
Students Receiving Title 1 Services	1	0	0	1										231	0	13	49	37	1132						
All Other Students	126	3	5	118	5	4	55	47	33	28	25	21	1141	12,586	6	39	33	21	1140						
504 Plan																									
Students with a 504 plan	5	0	0	5										599	4	38	36	22	1139						
All Other Students	122	3	5	114	5	4	53	46	32	28	24	21	1141	12,218	6	39	34	21	1140						

Level 4 = Proficient with Distinction Level 3 = Proficient Level 2 = Partially Proficient Level 1 = Substantially Below Proficient

Note: Some achievement level results have been left blank because fewer than ten (10) students were tested. **N** = Number

Maine
High School
Assessment

SCIENCE RESULTS

Test Date: May 2014
SAU: Demonstration District A

ACHIEVEMENT LEVELS: Achievement level definitions describe the quality of a student's responses on state-level assessments in relation to the science standards for achieving Maine's *Learning Results*.

Maine state-level assessments measure the knowledge and skills of students by sampling identified standards within science at the grade level assessed. Evidence includes responses to a combination of multiple-choice items and items requiring student-created responses in an "on demand" setting.

STUDENTS AT EACH ACHIEVEMENT LEVEL		School		SAU		State	
		N	%	N	%	N	%
Proficient with Distinction – The student's work demonstrates in-depth understanding of essential concepts in science, including the ability to make multiple connections among central ideas. The student's responses demonstrate the ability to synthesize information, analyze and solve difficult problems, and explain complex concepts using evidence and proper terminology to support and communicate logical conclusions. (Scaled Score 1162–1180)	2011–2012			4	3	650	5
	2012–2013			3	3	470	4
	2013–2014			**6**	**5**	**494**	**4**
	Cum. Average*			13	4	1,614	4
Proficient – The student's work demonstrates a general understanding of essential concepts in science, including the ability to make connections among central ideas. The student's responses demonstrate the ability to analyze and solve routine problems and explain central concepts with sufficient clarity and accuracy to demonstrate general understanding. (Scaled Score 1142–1160)	2011–2012			46	38	5,245	40
	2012–2013			36	32	4,782	37
	2013–2014			**54**	**45**	**5,090**	**40**
	Cum. Average*			136	39	15,117	39
Partially Proficient – The student's work demonstrates incomplete understanding of essential concepts in science and inconsistent connections among central ideas. The student's responses demonstrate some ability to analyze and solve problems but the quality of responses is inconsistent. Explanation of concepts may be incomplete or unclear. (Scaled Score 1134–1140)	2011–2012			31	26	3,413	26
	2012–2013			27	24	3,446	27
	2013–2014			**25**	**21**	**3,484**	**27**
	Cum. Average*			83	24	10,343	27
Substantially Below Proficient – The student's work demonstrates limited understanding of essential concepts in science and infrequent or inaccurate connections among central ideas. The student's responses demonstrate minimal ability to solve problems. Explanations are illogical, incomplete, or missing. There are many inaccuracies. (Scaled Score 1100–1132)	2011–2012			39	33	3,970	30
	2012–2013			47	42	4,105	32
	2013–2014			**35**	**29**	**3,693**	**29**
	Cum. Average*			121	34	11,768	30

Learning Results Content Strands	Number of Points Possible		Average Points Attained (Number and Percent)					
			School		SAU		State	
	N	%	N	%	N	%	N	%
Science Total Points	**56**	**100**			23.8	42.5	22.8	40.7
D. The Physical Setting	**34**	**61**			13.9	40.9	12.9	37.9
D1/D2 Space/Earth	**12**	**21**			5.1	42.5	4.6	38.3
D3/D4 Matter and Energy/Force and Motion	**22**	**39**			8.7	39.5	8.2	37.3
E. The Living Environment	**22**	**39**			10.0	45.5	9.9	45.0

The MHSA assesses students' science knowledge based on questions that measure the science accountability content strands highlighted in Maine's 2007 *Learning Results: Parameters for Essential Instruction*, which can be found at http://www.maine.gov/education/lres/pei/index.html.

Content Strand D. The Physical Setting
D1 - Universe and Solar System
D2 - Earth
D3 - Matter and Energy
D4 - Force and Motion

Content Strand E. The Living Environment
E1 - Biodiversity
E2 - Ecosystems
E3 - Cells
E4 - Heredity and Reproduction
E5 - Evolution

* Percentages are calculated by dividing the cumulative total of the number of students in the achievement level by the cumulative total of the number of students tested.

Maine
High School
Assessment

SCIENCE RESULTS
BY REPORTING SUBGROUPS

Test Date: May 2014
SAU: Demonstration District A

REPORTING CATEGORIES	SAU													State											
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
All Students	**127**	**2**	**5**	**120**	**6**	**5**	**54**	**45**	**25**	**21**	**35**	**29**	**1142**	**12,761**	**4**	**40**	**27**	**29**	**1141**						
Gender																									
Male	64	1	2	61	4	7	28	46	12	20	17	28	1143	6,594	5	41	25	28	1142						
Female	63	1	3	59	2	3	26	44	13	22	18	31	1141	6,167	2	38	30	30	1140						
Not Reported	0	0	0	0										0											
Primary Race/Ethnicity																									
Hispanic or Latino	1	0	0	1										185	1	35	31	33	1139						
Not Hispanic or Latino																									
American Indian or Alaskan Native	1	0	0	1										92	1	26	33	40	1138						
Asian	4	0	0	4										165	7	47	24	22	1144						
Black or African American	5	0	0	5										405	1	15	27	57	1134						
Native Hawaiian or Pacific Islander	1	0	0	1										14	0	50	21	29	1141						
White (non-Hispanic)	114	2	5	107	6	6	48	45	23	21	30	28	1142	11,774	4	41	27	28	1141						
Two or more races	1	0	0	1										126	3	42	28	27	1141						
LEP Status																									
Currently LEP student	2	0	0	2										247	0	4	18	78	1130						
Former LEP student - monitoring year 1	1	0	0	1										35	0	17	26	57	1135						
Former LEP student - monitoring year 2	1	0	0	1										59	0	20	47	32	1138						
All Other Students	123	2	5	116	6	5	53	46	25	22	32	28	1142	12,420	4	41	27	28	1141						
IEP																									
Students with an IEP	16	2	0	14	0	0	0	0	1	7	13	93	1130	1,679	1	12	19	69	1132						
All Other Students	111	0	5	106	6	6	54	51	24	23	22	21	1143	11,082	4	44	29	23	1142						
SES																									
Economically Disadvantaged Students	52	1	3	48	0	0	20	42	8	17	20	42	1138	4,581	1	28	29	42	1137						
All Other Students	75	1	2	72	6	8	34	47	17	24	15	21	1144	8,180	6	47	26	22	1143						
Migrant																									
Migrant Students	1	0	1	0										2											
All Other Students	126	2	4	120	6	5	54	45	25	21	35	29	1142	12,759	4	40	27	29	1141						
Title 1																									
Students Receiving Title 1 Services	1	0	0	1										227	<1	15	36	49	1135						
All Other Students	126	2	5	119	6	5	54	45	24	20	35	29	1142	12,534	4	40	27	29	1141						
504 Plan																									
Students with a 504 plan	5	0	0	5										590	2	42	27	29	1141						
All Other Students	122	2	5	115	6	5	51	44	25	22	33	29	1142	12,171	4	40	27	29	1141						

Level 4 = Proficient with Distinction   Level 3 = Proficient   Level 2 = Partially Proficient   Level 1 = Substantially Below Proficient

Note: Some achievement level results have been left blank because fewer than ten (10) students were tested.   **N** = Number

Maine
High School
Assessment

SCIENCE RESULTS
QUESTIONNAIRE ITEMS

Test Date: May 2014
SAU: Demonstration District A

QUESTIONNAIRE ITEMS	SAU										State											
	Students in Each Category	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Students in Each Category	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Students in Each Category	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	%	N	%	N	%	N	%	N	%		%	%	%	%	%		%	%	%	%	%	
How often do you make observations and collect data in science class?																						
A. a few times a week	36	2	5	22	51	8	19	11	26	1143	39	4	41	28	27	1141						
B. a few times a month	45	4	8	27	51	10	19	12	23	1144	40	5	46	27	23	1143						
C. once a month	12	0	0	3	21	5	36	6	43	1136	12	4	36	28	33	1140						
D. never or almost never	7										10	1	21	27	51	1135						
How do you feel about the following statement?																						
"My knowledge of science will be useful to me as an adult."																						
A. strongly agree	25	4	13	16	53	4	13	6	20	1145	23	11	54	21	14	1147						
B. agree	45	2	4	22	42	12	23	17	32	1142	48	3	42	28	27	1141						
C. disagree	24	0	0	15	54	6	21	7	25	1141	22	1	30	32	38	1138						
D. strongly disagree	6										7	1	17	27	55	1134						
What best describes your ninth grade science class?																						
A. earth/space science	48	2	4	31	54	9	16	15	26	1144	45	3	39	28	30	1141						
B. physical science	22	0	0	13	50	6	23	7	27	1141	23	3	39	30	29	1141						
C. engineering and physical science	4										3	3	35	25	36	1139						
D. mixture of physical science and life science	21	4	16	7	28	6	24	8	32	1142	23	5	45	26	25	1142						
E. physics	4										6	8	43	22	27	1143						
Do you think you would like to have a job that is related to SCIENCE?																						
A. No, this type of job is too hard.	6										6	<1	12	27	61	1133						
B. No, I'm not interested.	39	0	0	18	39	12	26	16	35	1139	41	1	33	30	35	1139						
C. I might be interested if I knew more about this type of job.	19	0	0	11	48	5	22	7	30	1141	19	2	40	30	28	1141						
D. Yes, I have some interest.	22	2	8	15	58	6	23	3	12	1146	18	5	52	25	18	1144						
E. Yes, I'm very interested.	14	4	25	9	56	1	6	2	13	1151	15	13	57	20	11	1148						
Which of the following best describes how you rate yourself as a student in science?																						
A. very good	11	3	23	7	54	1	8	2	15	1153	11	20	59	10	11	1152						
B. good	43	2	4	36	71	9	18	4	8	1146	43	3	53	27	17	1144						
C. fair	33	1	3	11	28	10	26	17	44	1137	38	<1	27	33	40	1137						
D. poor	13	0	0	0	0	5	33	10	67	1132	8	<1	11	28	61	1133						
How well do the questions that you have just been given on this MHSA test match what you have learned in school about science?																						
A. The questions on the test match what I have learned in science class.	10	1	8	9	75	2	17	0	0	1149	11	11	53	20	15	1147						
B. They match some of what I have learned.	57	4	6	34	51	14	21	15	22	1144	55	4	47	27	22	1143						
C. They match just a little of what I have learned.	28	1	3	11	33	9	27	12	36	1139	29	1	27	33	39	1137						
D. There is no match.	5										5	<1	10	20	70	1132						
Do you think you would like to have a job that is related to MATH?																						
A. No, this type of job is too hard.	8										8	1	23	29	47	1136						
B. No, I'm not interested.	44	1	2	26	50	13	25	12	23	1142	39	2	37	29	31	1140						
C. I might be interested if I knew more about this type of job.	25	2	7	11	37	9	30	8	27	1141	22	3	40	29	28	1141						
D. Yes, I have some interest.	17	2	10	10	50	2	10	6	30	1146	21	6	47	26	21	1144						
E. Yes, I'm very interested.	6										10	11	53	18	18	1147						

Level 4 = Proficient with Distinction Level 3 = Proficient Level 2 = Partially Proficient Level 1 = Substantially Below Proficient

August 2014

Maine Department of Education

2013-2014 School Year Reports

Dear School Board Members and School Personnel:

The Maine High School Assessment is the State's measure of student progress in achieving the State standards known as *Learning Results*. It consists of the SAT Reasoning Test™ (SAT) and a science test, and is administered to students in their third year of high school for state and federal purposes.

These Maine High School Assessment Summary Reports contain the results of your students' performance in critical reading, mathematics, writing, and science reported according to the academic standards described above and disaggregated by student and school characteristics. The MHSA achievement level standards for the critical reading, writing, mathematics, and science sections of the MHSA were determined by Maine educators with specific expertise within the content areas. This report, together with individual student and subject-specific student item-level reports, provides support for use in program evaluation and planning. All scores contained in these reports are included for Maine state and federal reporting purposes only. While scores from the SAT may also be used for college admission by most students, they may not be used for that purpose if a student received accommodations during the test administration that exceeded those made available by the College Board.

These results reflect scores based on SAT and science test questions that were taken by the nearly 14,000 publicly-funded students who were enrolled in their third year of high school across all Maine schools. The MHSA employs an assessment design that requires students to create an essay response to a writing prompt, generate answers to open-ended mathematics and science questions, and select answers to multiple-choice questions in all four disciplines. More information about the design, history, and use of the SAT can be found at: http://www.maine.gov/education/mhsa/index.htm.

I look forward to working with you in support of our continued efforts to improve the quality and effectiveness of the instructional opportunities designed to help all students achieve the high standards of the *Learning Results* and graduate from any Maine high school prepared for college, career, and citizenship.

Sincerely,

James E. Rier, Jr.
Commissioner of Education

Maine High School Assessment

High School Report

Test Date: May 2014

Code: DEMA-DEM3

SAU: Demonstration District A

School: Demonstration School 3

Contents of the Report

Due to small school size, this report contains only a summary of student participation to protect student confidentiality.

Topic ***Page***

Maine
High School
Assessment

SUMMARY OF STUDENT PARTICIPATION

Test Date: May 2014
SAU: Demonstration District A
School: Demonstration School 3

CATEGORY OF PARTICIPATION		Enrollment[1] during testing window						CONTENT AREA PARTICIPATION[2]																							
								Critical Reading						Mathematics						Writing						Science					
		School		SAU		State		School		SAU		State		School		SAU		State		School		SAU		State		School		SAU		State	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Total number of students		4	100	127	100	13574	100	4	100	121	96	13031	96	4	100	121	96	13039	96	4	100	120	96	13009	96	4	100	121	96	12952	95
Ethnicity	Hispanic or Latino	0	0	1	1	192	1	0	0	1	100	187	97	0	0	1	100	189	98	0	0	1	100	184	97	0	0	1	100	188	98
Not Hispanic or Latino	American Indian or Alaskan Native	0	0	1	1	103	1	0	0	1	100	93	90	0	0	1	100	93	90	0	0	1	100	95	92	0	0	1	100	95	93
	Asian	0	0	4	3	178	1	0	0	4	100	176	99	0	0	4	100	176	99	0	0	4	100	172	99	0	0	4	100	167	94
	Black or African American	0	0	5	4	442	3	0	0	5	100	423	96	0	0	5	100	426	97	0	0	4	100	401	96	0	0	5	100	416	94
	Native Hawaiian or Pacific Islander	0	0	1	1	14	<1	0	0	1	100	14	100	0	0	1	100	14	100	0	0	1	100	14	100	0	0	1	100	14	100
	White	4	100	114	90	12512	92	4	100	108	96	12011	96	4	100	108	96	12014	96	4	100	108	96	12017	96	4	100	108	96	11945	96
	Two or more races	0	0	1	1	133	1	0	0	1	100	127	96	0	0	1	100	127	96	0	0	1	100	126	96	0	0	1	100	127	95
Identified disability		1	25	16	13	2051	15	1	100	15	94	1852	91	1	100	15	94	1853	91	1	100	15	94	1859	91	1	100	15	100	1870	91
Current LEP		0	0	2	2	285	2	0	0	2	100	270	95	0	0	2	100	271	95	0	0	1	100	232	94	0	0	2	100	258	91
Economically disadvantaged		3	75	52	41	4999	37	3	100	48	94	4688	94	3	100	48	94	4699	94	3	100	47	94	4675	94	3	100	48	94	4683	94
Migrant		0	0	1	1	4	<1	0	0	1	100	4	100	0	0	1	100	4	100	0	0	1	100	4	100	0	0	0	0	2	50

MODE OF PARTICIPATION[3]	Critical Reading						Mathematics						Writing						Science					
	School		SAU		State		School		SAU		State		School		SAU		State		School		SAU		State	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Participation without accommodations	4	100	108	85	11522	85	4	100	108	85	11514	85	4	100	108	85	11504	85	4	100	109	86	11512	85
Identified disability (IEP)	1	25	6	6	765	7	1	25	6	6	761	7	1	25	6	6	765	7	1	25	7	6	821	7
LEP	0	0	1	1	188	2	0	0	1	1	188	2	0	0	1	1	169	1	0	0	1	1	183	2
Participation with accommodations	0	0	11	9	1299	10	0	0	12	9	1331	10	0	0	11	9	1313	10	0	0	11	9	1249	9
Identified disability (IEP)	0	0	8	73	892	69	0	0	8	67	898	67	0	0	8	73	902	69	0	0	7	64	858	69
LEP	0	0	0	0	56	4	0	0	1	8	72	5	0	0	0	0	52	4	0	0	1	9	64	5
Participation through alternate assessment (PAAP)	0	0	1	1	195	1	0	0	1	1	194	1	0	0	1	1	192	1	0	0	1	1	191	1
Identified disability (IEP)	0	0	1	100	195	100	0	0	1	100	194	100	0	0	1	100	192	100	0	0	1	100	191	100
LEP	0	0	0	0	11	6	0	0	0	0	11	6	0	0	0	0	11	6	0	0	0	0	11	6
Approved non-participation in reading – 1st year LEP	0	0	1	1	15	<1																		
Approved non-participation – special consideration	0	0	1	1	20	<1	0	0	1	1	20	<1	0	0	2	2	58	<1	0	0	1	1	10	<1
Non-participation – other	0	0	5	4	523	4	0	0	5	4	515	4	0	0	5	4	507	4	0	0	5	4	612	5

[1] Percents are the percentage of students enrolled in each participation category.

[2] Percents are the percentage of students, including those who participated through alternate assessment (PAAP), who participated in the content area.

[3] Percents are the percentage of students in each content area by mode.

MeCAS
Maine's Comprehensive
Assessment System

State Summary

2013-2014 Science

State: Maine

Science	Enrolled	Not Tested Approved	Not Tested Other	Tested	Achievement Level								
	N	N	N	N	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score
					N	%	N	%	N	%	N	%	
Maine	**40,790**	**706**	**870**	**39,214**	**4,873**	**12**	**18,723**	**48**	**9,867**	**25**	**5,751**	**15**	
Grade 5	13,296	215	101	12,980	1,301	10	6,859	53	3,783	29	1,037	8	546
Grade 8	13,920	290	157	13,473	3,078	23	6,774	50	2,600	19	1,021	8	850
High School	13,574	201	612	12,761	494	4	5,090	40	3,484	27	3,693	29	1141

Level 4 = Proficient with Distinction; Level 3 = Proficient; Level 2 = Partially Proficient; Level 1 = Substantially Below Proficient

APPENDIX O—INTERACTIVE REPORTS

Maine
High School
Assessment

C O N F I D E N T I A L

Science Item Analysis Report - Spring 2014

High School

Date:	8/26/2014 11:02:51 AM
Code:	DEMA-DEM1
Group Size:	69
SAU:	Demonstration District A
School:	Demonstration School 1

Page: 1 of 3

	Released Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Content Strand Points Earned				Total Points Earned	Scaled Score	Achievement Level
	Content Strand	D1	D1	E1	D3	E3	E4	E1	D4	E1	E4	E3	D4	D3	D1	E4	D3	D3	E3	D4	D2	E5	D4							
	Depth of Knowledge Code	1	1	2	2	2	1	2	2	2	2	1	2	2	2	1	3	3	2	2	3	3	2	D. The Physical Setting			E. The Living Environment			
	Item Type	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	CR	CR							
	Answer Key	A	D	A	B	C	A	B	B	A	A	A	D	C	B	A	B	C	B	C	A			D. Total	D1/D2	D3/D4	E. Total			
Name/MEDMS ID	Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	4	34	12	22	22	56		
Altvater, Diahnie	D11100008		+	+	D		C	+	+	+	+	+	+	+	C	C	C		+		C	3	1	16.33	5.00	11.33	14.00	30.33	1146	3
Anderson, Christia C	D11100043	+	+	+	C	+	+	+	+	+	+	D	+	+	+	D	C	B	+	D	+	4	1	18.33	8.67	9.67	17.00	35.33	1152	3
Anderson, Kati A	D11100094	B	+	D	+	+	+	+	+	B	B	B	+	+	+	B	C	D	+	D	C	2	1	8.67	4.33	4.33	8.33	17.00	1136	2
Andrews, Nicholas E	D11100012	+	+	+	+	D	+	+	+	+	B	+	+	+	+	+	+	+	+	A	+	4	3	27.00	10.00	17.00	16.00	43.00	1160	3
Bagley, Jordan L	D11100038																													DNP
Beckham, Nicole A	D11100050	C	+	+	+		+	+	+	D	B	+	+	B	+	+	C	D	+	B	C	2	0	4.00	0.00	4.00	10.33	14.33	1132	1
Bergstrom, Robert A	D11100002	B	+	+	D	+	+	+	+	+	+	C	+	+	+	+	D	D	+	+	C	1	1	18.33	4.67	13.67	8.00	26.33	1144	3
Berosik, Tana A	D11100033	D	+	+	+	B	+	D	+	+	+	+	A	+	C	C	C	D	+	+	+	2	2	14.33	3.67	10.67	10.00	24.33	1140	2
Billadodubie, Damien R	D11100060	B	+	+	+	A	B	A	+	D	C	+	A	+	C	B	C	B	C	A	+	1	1	9.33	3.67	5.67	2.67	12.00	1132	1
Bishop, Morgan R	D11100020	B	+	+	+	+	+	C	+	C	+	+	+	B	D	D	C	D	+	A	C	B	1	4.67	1.67	3.00	8.67	13.33	1132	1
Blair, Brandon S	D11100016	C	A	+	D	A	B	A	A	B	B	+	C	B	D	D	+	+	A	B	C	B	B	-2.00	-2.67	0.67	0.67	-1.33	1120	1
Blaskovich, Kiley A	D11100120		+		+	+	+	A	+	B	+	D	A		+	B	C	D	C	+	C	2	2	7.67	1.33	6.33	6.67	14.33	1132	1
Brooks, Colton R	D11100105		+	+	D		+	+	+	+	+		+		+		D	+	+		+	1	0	15.33	7.67	7.67	10.00	25.33	1142	3
Brown, Shannon N	D11100067	C		+		+	+	D	+	+	+	D	+	+	C	B	+		+	+	C	2	2	18.67	3.67	15.00	13.00	31.67	1148	3
Curtsinger, Mark G	D11100088	+	+	+	+	B	+	+	+	+	+	+	+	+	+	D	C	A	+	+	+	2	0	20.33	9.00	11.33	11.67	32.00	1148	3
D' Agostino, Jordan M	D11100079	B	B	+	+	+	C	D	+	+	+	B	+	+	+	B	+	B	+	A	C	3	1	15.67	3.33	12.33	9.67	25.33	1142	3
Dearmond, Kayla	D11100101		+	+		+		+	+	+	+	D	+	+	+	C	D		+	+	C	3	2	20.67	8.33	12.33	13.00	33.67	1150	3
Degidio, Vito	D11100011	+	+	+	D	+	+	+	+	+	+	+	+	+	+	+	+	+	+	D	+	2	2	22.33	7.67	14.67	13.33	35.67	1152	3
Doyle, Cameron B	D11100118	+	C	+	A	A	B	+	D	+	+	D	B	+	+	D	C	B	A	+	C	3	1	9.67	4.00	5.67	4.67	14.33	1132	1
Ehrlich, Daniel D	D11100019	D	+	+	+	+	+	+	+	+	+	C	C	B	+	C	C	D	+	+	+	1	2	14.00	7.33	6.67	8.00	22.00	1140	2
Einstein, Jaclynn E	D11100017	+	+	D	+	+	D	A	+	+	+	+	+	B	+	B	+	+	+	+	C	3	2	13.33	1.33	12.00	10.33	23.67	1140	2
Elder, Timothy	D11100032	+	+	C	+	B	+	+	+	+	+	+	+	+	+	+	+	D	+	+	+	4	2	25.67	9.67	16.00	15.67	41.33	1158	3
Ellis, Charlott	D11100102																													ASC
Fehrnstrom, Jason D	D11100126	C	+	+	+	+	+	C	D	C	D	B	+	+	C	B	D	+	+	A	C	1	1	7.00	2.67	4.33	5.67	12.67	1132	1
Firestone, Shirley L	D11100061	B	+	+	D	B	+	+	D	+	+	B	+	B	+	B	+	D	D	A	C	3	1	11.67	6.00	5.67	7.67	19.33	1138	2

Maine High School Assessment

C O N F I D E N T I A L

Science Item Analysis Report - Spring 2014

High School

Date: 8/26/2014 11:02:51 AM
Code: DEMA-DEM1
Group Size: 69
SAU: Demonstration District A
School: Demonstration School 1

	Released Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Content Strand Points Earned				Total Points Earned	Scaled Score	Achievement Level
	Content Strand	D1	D1	E1	D3	E3	E4	E1	D4	E1	E4	E3	D4	D3	D1	E4	D3	D3	E3	D4	D2	E5	D4							
	Depth of Knowledge Code	1	1	2	2	2	1	2	2	2	2	1	2	2	2	1	3	3	2	2	3	3	2	D. The Physical Setting			E. The Living Environment			
	Item Type	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	CR	CR							
	Answer Key	A	D	A	B	C	A	B	B	A	A	A	D	C	B	A	B	C	B	C	A			D. Total	D1/D2	D3/D4	E. Total			
Name/MEDMS ID	Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	4	34	12	22	22	56		
Fortin, Brandon E	D11100014																													DNP
Frazier, Rebecca	D11100113	C	+	+	+	+	+	+	+	+	+	D	A	D	+	B	C	D	C		+	1	1	6.33	2.67	3.67	8.33	14.67	1132	1
Gonzalez, Rylie A	D11100025	+	+		+			+	+	+	+	+	B	+	+			+	+		+	3	2	20.67	11.00	9.67	12.67	33.33	1150	3
Gonzalezaguilar, Luis V	D11100116	+	B	+	D	+	C	+	+	+	+	+	+	+	+	D	+	B	+	B	+	3	2	19.00	9.67	9.33	16.00	35.00	1152	3
Hernandez, Thania P	D11100083	B	A	+		A	+	D	A	D	B	+	+	+	+	B	D	+	+	D	C	1	0	4.67	2.00	2.67	5.67	10.33	1130	1
Hess, Zachary	D11100093	D	+	+	D	D	+	C	+	+	+	+	+	+	+	D	+	B	+	+	+	2	1	17.00	8.67	8.33	6.67	23.67	1140	2
Jarrard, Eric D	D11100054	+	A	+	D	D	+	+	+	+	+	C	+	+	C	C	+	B	D	A	D	1	0	7.33	2.67	4.67	8.00	15.33	1132	1
Johnson, Laura E	D11100065	C	+	+	+	A	+	+	+	+	+	+	+	+	C	+	D	+	+	+	+	1	1	18.67	6.33	12.33	14.00	32.67	1150	3
Jones, Eric P	D11100053	C	+	+	D	B	D	+	+	+	+	B	+	B	+	D	D	D	+	B	+	2	1	17.33	6.33	11.00	8.67	26.00	1144	3
Joyce, Jillian M	D11100082	D	B	+	D	D	B	+	+	C	B	D	C	+	+	B	+	D	C	B	+	0	0	6.00	2.67	3.33	0.67	6.67	1128	1
Kadrmas, Caleb C	D11100064	B	+	+	D	D	+	+	D	+	+	B	+	+	+	+	C	D	+	D	+	2	3	15.33	6.33	9.00	9.33	24.67	1142	3
Kalloch, Casey	D11100004	B	A	+	D	A	B	D	A	+	C	D	+	+	+	B	C	B	A	B	C	1	B	2.00	0.00	2.00	2.33	4.33	1126	1
Keith, Christopher E	D11100111	D	+	+	D		+	+	+	D	+	B	A	B	C	C	A	D	D	A	C	1	0	7.00	1.00	6.00	3.33	10.33	1130	1
Kelso, Kyren F	D11100027	+	+	+	D	D	B	+	+	D	B	+	+	+	+	D	C	D	+	D	C		0			2.00		11.67	1132	1
Madden, Alicia M	D11100117	B	+	+	C	D	D	+	D	+	+	+	+	+	+	B	A	+	+	+	+	3	1	21.33	6.33	15.00	14.67	36.00	1152	3
Malicdem, Raquelle M	D11100110	B	B	+		A		D	+	+	D	+		D	+	B	A	B		D	C	1	0	2.67	1.33	1.33	3.00	5.67	1128	1
Malik, Doren	D11100058																													DNP
Maron, Kasey	D11100087	+	+	+	+	+	+	+	+	+	B	+	+	+	+	B		+	+		+	3	2	26.33	10.67	15.67	9.33	35.67	1152	3
McMillan, Shayla	D11100030	+	+	+	+	A	D	C	+	D	+	D	A	A	D	C	C	B	C	A	+	0	0	2.00	2.67	-0.67	-0.67	1.33	1122	1
McTaggart, Riley D	D11100045	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	D	+	+		+	3	1	22.00	9.33	12.67	17.00	39.00	1156	3
Mendoza, Cristian E	D11100046	B	+	+			D	+	+	+	+	D	+	B	+	B	+		+	+	C	1	2	11.00	2.67	8.33	7.00	18.00	1136	2
Messinajr, John J	D11100091	B	+	+	D	A	D	+	D	D	+	+	B	+	C	B	+	+	D	+	+	2	1	13.00	4.67	8.33	5.33	18.33	1138	2
Naylor, Clare	D11100034	+	+	+	D	+	+	+	+	+	+	+	+	+	+		+	B	+		C	4	2	23.33	8.33	15.00	21.00	44.33	1164	4
Neri, Andy	D11100068	+	+	+	D	+	+	+	+	+	+	+	+	+	+	+	+	A	+	+	C	3	2	18.67	9.33	9.33	16.67	35.33	1152	3
Olsen, Makenna N	D11100125	B	+	+	A	B	+	D	+	D	B	+	+	B	A	B	C	+	A	+	C	1	1	7.00	0.00	7.00	4.33	11.33	1132	1

Maine
High School
Assessment

C O N F I D E N T I A L

Science Item Analysis Report - Spring 2014

High School

Date: 8/26/2014 11:02:51 AM
Code: DEMA-DEM1
Group Size: 69
SAU: Demonstration District A
School: Demonstration School 1

	Released Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Content Strand Points Earned				Total Points Earned	Scaled Score	Achievement Level
	Content Strand	D1	D1	E1	D3	E3	E4	E1	D4	E1	E4	E3	D4	D3	D1	E4	D3	D3	E3	D4	D2	E5	D4							
	Depth of Knowledge Code	1	1	2	2	2	1	2	2	2	2	1	2	2	2	1	3	3	2	2	3	3	2	D. The Physical Setting			E. The Living Environment			
	Item Type	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	CR	CR							
	Answer Key	A	D	A	B	C	A	B	B	A	A	A	D	C	B	A	B	C	B	C	A			D. Total	D1/D2	D3/D4	E. Total			
Name/MEDMS ID	Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	4	34	12	22	22	56		
Ortiz, Walter J	D11100009	+	+	+	+	D	+	C	D	+	+	D	C	+	+	B	D	+	+	+	C	4	2	17.33	4.00	13.33	11.33	28.67	1146	3
Ruiz, Blanca	D11100073	B	+	+	D	A	+	D	D	+	B	D	+	+	+	B	C	B	+	+	C	2	2	10.67	1.33	9.33	5.33	16.00	1136	2
Ruiz, Mario J	D11100036	+	+	+	+	+		+	+	+	+		+	+	+	D	+	B	A	+	C	3	1	26.00	10.67	15.33	11.00	37.00	1154	3
Salazar, Juan	D11100090	+	+	+	+	A	+	+	+	+	+	+	+	+	+	D	+	A	+	+	+	3	3	29.00	12.00	17.00	16.33	45.33	1164	4
Schuler, Sarah	D11100104																													DNP
Scremin, Matthew R	D11100122	+	+	+	+	+	C	+	+	+	+	+	A	+	+	+	C	B	+	D	C	2	3	23.00	8.67	14.33	14.33	37.33	1154	3
Simpatico, Daniella L	D11100022	B	B	+	+	A	C	+	+	D	+	C	+	+	+	D	C	D	+	B	C	2	1	9.67	2.67	7.00	6.33	16.00	1136	2
Sivertson, Tegan E	D11100078	+	A	+	D	+	+	D	+	+	B	+	+	+	+	B	D	+	+	D	+	1	1	18.33	7.33	11.00	10.67	29.00	1146	3
Talluri, Harika W	D11100031		+		D	A	+	D	+	+	C	B	+		+	B		D	+	+	C	2	1	12.67	2.67	10.00	5.33	18.00	1136	2
Taylor, Isaiah T	D11100047	+	+	+	+	+		+	+	+	+	+	+	+	+	C	C	+	+	D	+	4	B	24.00	12.00	12.00	17.67	41.67	1160	3
Tharpe, Lindsey C	D11100010	+	+	+		+	D	+	+	+	+	+	+	+	C	+	+	+	+		C	3	3	24.67	9.33	15.33	19.67	44.33	1164	4
Turgeon, Brian	D11100092	C	+	+	D	B	+	D	+	+	+	+	A	B	+	B	C	+	C	+	+	1	0	17.00	7.00	10.00	10.33	27.33	1144	3
Tyler, Autumn	D11100100		+	+	D	+	+	+	+	+	B	D	+	+	+	C	C	A	+	A		2	1	7.00	3.67	3.33	12.00	19.00	1138	2
Vorhies, Kyra C	D11100071	D	+	+	D	+	+	+	D	+	+	+	+	+	+	+	D	+	+	+	C	1	1	12.33	4.00	8.33	11.67	24.00	1140	2
Washburn, Matthew J	D11100119	+	+	+		+	B	+	+	+	+	+	+	+	+	+	+	A	+	+	B	3	1	17.33	6.67	10.67	17.67	35.00	1152	3
Williams, Haylee K	D11100121	B	+	+	D	+	+	D		+	+	+	+	+	+	B	D	B	+		C	3	2	14.00	7.00	7.00	14.33	28.33	1146	3
Wilson, Alexandria L	D11100018	B	+	+	C	B	D	D	+	+	+	+	+	B	+	C	C	B	+	D	C	1	0	7.67	4.00	3.67	7.67	15.33	1132	1
Wipf, Mike A	D11100052	B	+	+	+	+		+	+	+	+	D	+	B	+	B		+	+	D		1	1	11.00	3.33	7.67	8.33	19.33	1138	2
Wolfe, Colt	D11100006	B	+	+	D	B	B	D	+	B		D	C	+	D	B	C	D	+	+	C	1	B	8.00	1.33	6.67	2.33	10.33	1130	1
Released Item		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22							
Percent Correct/Avg. Score: Group		37	81	90	41	44	59	63	79	76	73	54	73	70	76	21	32	33	75	41	41	2.00	1.20	14.30	5.30	9.00	9.90			
Percent Correct/Avg. Score: School		37	81	90	41	44	59	63	79	76	73	54	73	70	76	21	32	33	75	41	41	2.00	1.20	14.30	5.30	9.00	9.90			
Percent Correct/Avg. Score: SAU		43	74	90	43	43	60	60	73	78	72	55	74	69	73	25	29	35	80	39	43	2.00	1.30	13.90	5.10	8.70	10.00			
Percent Correct/Avg. Score: State		42	67	88	42	45	62	57	75	78	72	54	70	69	74	25	33	29	78	45	41	1.90	1.20	12.90	4.60	8.20	9.90			

Maine
High School
Assessment

Item Analysis Report Legend

SAT Critical Reading, Mathematics, and Writing

MHSA Science

May 2014

Released Item: This report provides data on items that are being released, which represent a percentage of the common items taken by each student. For the three portions of the SAT, Critical Reading, Mathematics, and Writing, 100% of the items are being released. For the Science portion of the test, 50% of the items are being released. The MHSA consists of common items and field test items but only the common items, both released and non-released, are used to calculate total MHSA scores.

Section: This is the section of the SAT released test form in which the item appears.

Content Strand: For Science, the letter indicates the content standard with which an item is aligned as outlined in Maine's 2007 ***Learning Results: Parameters for Essential Instruction*** for grades 9–12. For science the performance indicator is also displayed. For SAT critical reading and mathematics, the content strands align with NECAP.

Depth of Knowledge Code: This number indicates the Depth of Knowledge to which the item is coded for Science only.

SAT Writing Category: The letters identify the SAT writing categories including the writing essay (ES).

Item Type: This indicates whether the question is multiple-choice (MC), a student-generated response (SG – SAT mathematics only), constructed-response (CR – science only), or essay (ES - writing only).

Answer Key: This is the correct letter response for the multiple-choice questions.

Possible Points: The number under each released item number indicates the maximum number of points that could be earned for the question. All multiple-choice questions are "formula scored" to compensate for guessing. The formula scoring process assigns 1 point to a correct answer, 0 points to an unanswered question, and deducts a fraction of a point for an incorrect answer. The fraction deducted for an incorrect mathematics, reading, or writing multiple-choice question is 1/4 of a point as all questions have 4 incorrect answer choices. The fraction deducted for an incorrect science multiple-choice question is 1/3 of a point as all science questions have 3 incorrect answer choices. Formula scoring is not applied to the science constructed-response questions; rather, these are scored using a 4-point rubric and no deductions are made for incorrect responses.

Name/ICSE ID: Each student's name and ICSE identification number are listed, followed by data for each released item on the test. For multiple choice responses, a plus sign (+) indicates a correct response, and a letter indicates the incorrect response chosen. A space indicates either no selection or more than one selection was made by the student. For science constructed responses or mathematics student generated responses, a number indicates the points earned, and the letter B indicates a blank response.

Total Points Earned: This column shows the total number of points earned on all items. For writing this total includes points earned on the SAT essay. **NRF** in this column means "Not Released Form," which appears for a student who took the SAT during any administration other than on May 3, 2014.

Scaled Score: This column shows the scaled score in a range from 1100–1180 that corresponds to the points earned.

Achievement Level: This column shows the achievement level into which the student's scores fall: **4** = Proficient with Distinction, **3** = Proficient, **2** = Partially Proficient, and **1** = Substantially Below Proficient. **DNP** indicates that the student did not participate due to a non-approved reason. In reading only, **LEP** indicates that the student did not participate due to first year enrollment in a United States school and participation in the WIDA ACCESS for ELLs®. **ALT** indicates that the student participated through a personalized alternate assessment portfolio (PAAP). **HLD** indicates that the student's scores are on hold as a result of some unresolved issue that occurred during the registration and/or administration of the SAT portion of the MHSA. **ASC** indicates that the student did not participate due to a state approved special consideration.

Percent Correct: Percent correct refers to the percent of students who answered a multiple-choice item correctly. Avg. score refers to the average of the number of points awarded to all students who attempted that constructed-response item (science only). These are listed by school, district, and state.

Rev. July 2014

Maine
High School
Assessment

Achievement Level Summary

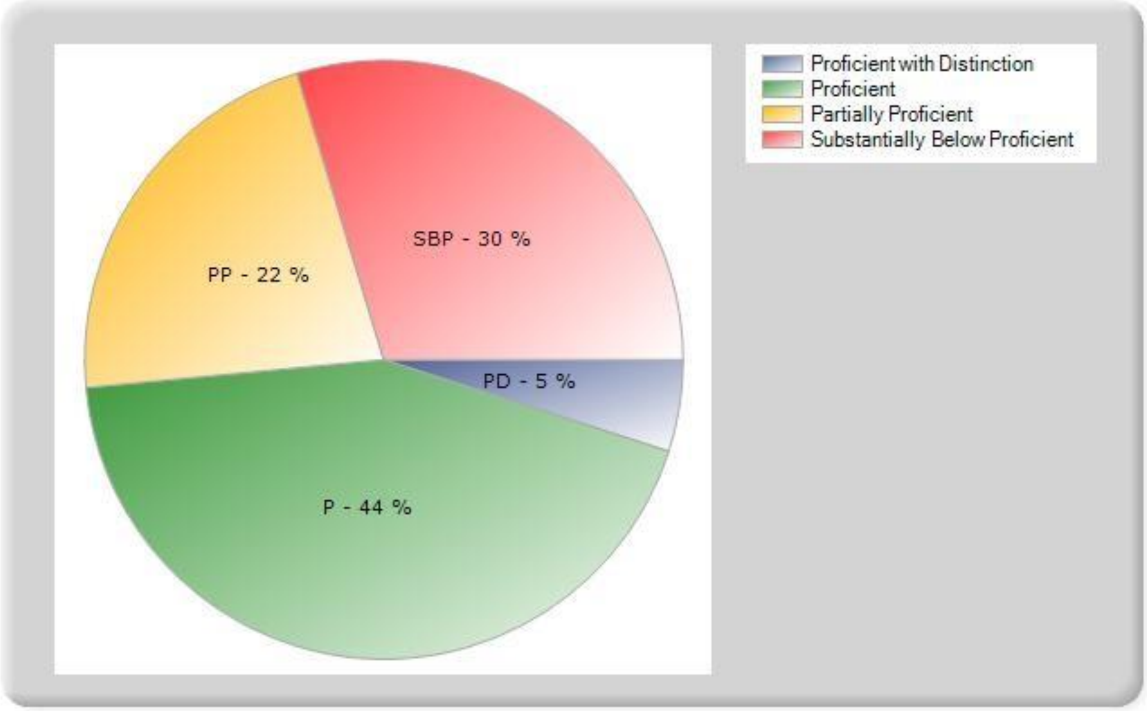
SAU: Demonstration District A

School: Demonstration School 1

Grade: 11

Date: 8/26/2014 11:05:37 AM

Science



Achievement Level	Count	Percentage %*
Proficient with Distinction	3	5
Proficient	28	44
Partially Proficient	14	22
Substantially Below Proficient	19	30

*Percentages may not total exactly 100% due to applied rounding.

Maine
High School
Assessment

Science Released Items Summary Data

SAU: Demonstration District A

School: Demonstration School 1

Grade: 11

Date: 8/26/2014 11:07:11 AM

Multiple Choice

Released Item	Content Strand	Correct (#)	A (#)	B (#)	C (#)	D (#)	IR (#)	Correct Response
1	D1	23	23	20	8	6	6	A
2	D1	51	5	5	1	51	1	D
3	E1	57	57	0	1	2	3	A
4	D3	26	2	26	3	25	7	B
5	E3	28	13	9	28	7	6	C
6	E4	37	37	7	5	8	6	A
7	E1	40	4	40	5	14	0	B
8	D4	50	3	50	0	9	1	B
9	E1	48	48	4	3	8	0	A
10	E4	46	46	11	3	2	1	A
11	E3	34	34	8	4	15	2	A
12	D4	46	8	3	5	46	1	D
13	D3	44	1	13	44	2	3	C
14	D1	48	1	48	10	4	0	B
15	E4	13	13	26	10	11	3	A
16	D3	20	3	20	24	12	4	B
17	D3	21	5	16	21	17	4	C
18	E3	47	5	47	6	4	1	B
19	D4	26	10	7	26	11	9	C
20	D2	26	26	1	33	1	2	A

Constructed Response

Released Item	Content Strand	Point Value	Average Score
21	E5	4	2.0
22	D4	4	1.2

CONFIDENTIAL

Longitudinal Data Report

Student Name
Andy Neri

Year	Enrolled Grade	School Name	Administration	Test Name	Content Area	Score	Achievement Level
0910	11	Demonstration School 1	MHSA 2010	Grade 11 Mathematics	mat	1154	Proficient
0910	11	Demonstration School 1	MHSA 2010	Grade 11 Reading	rea	1160	Proficient
0910	11	Demonstration School 1	MHSA 2010	Grade 11 Science	sci	1152	Proficient
0910	11	Demonstration School 1	MHSA 2010	Grade 11 Writing	wri	1166	Proficient with Distinction
1011	11	Demonstration School 2	MHSA 2011	Grade 11 Mathematics	mat	1138	Partially Proficient
1011	11	Demonstration School 2	MHSA 2011	Grade 11 Reading	rea	1148	Proficient
1011	11	Demonstration School 2	MHSA 2011	Grade 11 Science	sci	1144	Proficient
1011	11	Demonstration School 2	MHSA 2011	Grade 11 Writing	wri	1144	Proficient
1112	11	Demonstration School 2	MHSA 2012	Grade 11 Mathematics	mat	1164	Proficient with Distinction
1112	11	Demonstration School 2	MHSA 2012	Grade 11 Reading	rea	1158	Proficient
1112	11	Demonstration School 2	MHSA 2012	Grade 11 Science	sci	1148	Proficient
1112	11	Demonstration School 2	MHSA 2012	Grade 11 Writing	wri	1168	Proficient with Distinction
1213	11	Demonstration School 1	MHSA 2013	Grade 11 Mathematics	mat	1140	Partially Proficient
1213	11	Demonstration School 1	MHSA 2013	Grade 11 Reading	rea	1154	Proficient
1213	11	Demonstration School 1	MHSA 2013	Grade 11 Science	sci	1140	Partially Proficient
1213	11	Demonstration School 1	MHSA 2013	Grade 11 Writing	wri	1142	Proficient
1314	11	Demonstration School 1	MHSA 2014	Grade 11 Mathematics	mat	1144	Proficient
1314	11	Demonstration School 1	MHSA 2014	Grade 11 Reading	rea	1150	Proficient
1314	11	Demonstration School 1	MHSA 2014	Grade 11 Science	sci	1152	Proficient
1314	11	Demonstration School 1	MHSA 2014	Grade 11 Writing	wri	1146	Proficient

Note: This report returns as many years of NECAP data as are available for this student beginning with 08-09.

APPENDIX P—DECISION RULES

Analysis and Reporting Decision Rules
Maine High School Assessment
Spring 13-14Administration

This document details rules for analysis and reporting. The final student level data set used for analysis and reporting is described in the "Data Processing Specifications." This document is considered a draft until the Maine State Department of Education (DOE) signs off. If there are rules that need to be added or modified after said sign-off, DOE sign off will be obtained for each rule. Details of these additions and modifications will be in the Addendum section.

I. General Information

A. *Test administered:*

Grade	Subject	Items Included in Raw Score	Data Used for MHSA Scaled Scores
HS	Mathematics	*NA*	SAT Scaled Score
HS	Critical Reading	*NA*	SAT Scaled Score
HS	Writing	*NA*	SAT Scaled Score
HS	Science	Common	Science unrounded raw score

B. *Reports Produced*:

1. Individual Student Report (ISR) (printed/online)
2. Student Labels (printed)
3. Item Analysis Report (by subject) (online)

   This document specifies the data requirements for interactive reporting
4. Summary Report Package (online)
   a Reporting Levels: School, SAU and State
   b Package contains:
   - Summary of Scores
   - Summary of Student Participation
   - Results (by subject)
5. One Page Summary

   Report Specifications are in MEA Science Decision Rules. Rules for calculating Not Tested – State Approved and Not Tested – Other can be found in Section III.

C. *Files Produced:*

1. State Student Raw Data Files
   (With names and without names, With Science and without Science)
2. State Student Scored Data Files
   (With names and without names)
3. SAU/School Student Results Files *(by subject)*
4. Accountability Student Results - Specifications are in HS Accountability Decision Rules
5. Press Release

   (School, SAU)
6. State Accommodation Frequency Report
7. Top 50 HS Students
8. State Standard Deviations & Average Scaled Scores
   (by subject)
9. Minimally Statistically Significant Differences for Scaled Scores (*by subject)*
10. Standard Error of Measurement (Science only)
11. Scaled Score Lookup (Science only)
12. Score Range (Science only)
13. MHSA School List for Mailing(For Program Management)
14. State Student Questionnaire
15. Department Chair and Principal Questionnaire Raw Data
16. Department Chair/Principal Questionnaire Frequency Distribution

D. *School Type:*

SchType	Source: ICORE SubTypeID	Description
‘PUB’	1	Public
‘PSP’	19	Public Special Purpose
‘PSE’	15	Public Special Ed
‘BIG’	6	Private with 60% or more Publicly Funded (Big 11)
‘PSN’	23	Private Special Purpose
‘CHA’	11	Charter School

School Type impact on Data Analysis and Reporting		
Level	**Impact on Analysis**	**Impact on Reporting**
Student	n/a	Report students based on testing discode and schcode. SAU data will be blank for students tested at BIG or PSN schools. Always print tested year state data.
School	Do not exclude any students based on school type using testing school code for aggregations	Generate a report for each school with at least one student enrolled using the tested school aggregate denominator. SAU data will be blank for BIG and PSN schools. Always print tested year state data.
SAU	For BIG and PSN schools, aggregate using the sending SAU. If BIG or PSN student does not have a sending SAU, do not include in aggregations.	Generate a report for each SAU with at least one student enrolled using the tested SAU aggregate denominator. Always report tested year state data.
State	Include all students.	Always report testing year state data.

E. *Stustatus:*

StuStatus	**Description**
1	Homeschooled
2	Privately Funded
3	Exchange Student
4	Excluded State – removed before analysis
0	Publicly Funded

StuStatus impact on Data Analysis and Reporting		
Level	**Impact on Analysis**	**Impact on Reporting**
Student	n/a	School and SAU data will be blank for students with a StuStatus value of 1. Always print tested year state data. For StuStatus values of 1 School name is 'Home Schooled' and SAU name is the name of the student's reported SAU.
School	Exclude all students with a StuStatus value of 1, 2 or 3.	Students with a StuStatus value of 1, 2 or 3 are excluded from Interactive Reporting.
SAU	Exclude all students with a StuStatus value of 1, 2 or 3.	n/a
State	Exclude all students with a StuStatus value of 1, 2, 3.	n/a.

F. *Other Information*

1. 3rd year High School (HS) students are expected to test the SAT and MHSA science.
2. A non-public SAU code is a SAU associated with a school that is type BIG or PSN. Non-public testing sending SAU codes will be ignored.
3. Only students with a school type of BIG or PSN are allowed to have a sending SAU code. Sending SAU codes will be blanked for any other school type.
4. SAT Hold Students

    - Students whose SAT scores are on hold by the College Board will be reported as 'HLD' on the roster for math, reading, and writing. The students will not be included in participation and performance aggregations for those content areas.

    - If the student participated on the Science test, he/she will be reported as defined by the rules in this document.

- If the College Board releases the student’s results, Measured Progress will do a rerun so that the students will get results in all content areas prior to HA Accountability reporting.

5. Home Schooled Students

   a A student is identified as Home School based on the Student demographic file data (Stustatus = ‘1’).

   b Home schooled students will only be included in all content areas of reporting if the entire MHSA is completed.

   c If only Science is submitted, the student will be reported only for Science.

   d If only the SAT is submitted, the student will not be included in MHSA reporting.

   e Home schooled students only appear on ISR reports.

6. Student Demographic File Linking

   a If a student is links to the student demographic file, all demographic data of record are pulled from the student demographic file.

   b If the student does not link to Student demographic file the student will not be reported.

7. Inactive Students (Active does not = ‘2’)

   a If a student is not active in the student demographic file and completes the SAT portion of the MHSA, the student will be reported as defined by the rules in this document.

   b If a student is not active in the student demographic file and only completes the Science portion of the MHSA, the results will be suppressed and the student will not be reported.

8. Grade 11 students who are not marked as Active=’2’ in the Student Demographic file will receive a parent letter. They are not included in interactive reporting and any aggregations.

9. Students are removed prior to analysis if any of the following conditions are true:

   a Student does not have a valid student ID in the student demographic file id.

   b Student grade is not 11 and student is not marked as Active=’2’ in the Student Demographic file

   c Stustatus = ‘4’ in the student demographic file.

   d Student is inactive in the student demographic file (Active = ‘0’) and only the Science portion of the MHSA is submitted.

   e Student is home schooled (Stustatus= ‘1) and only the SAT portion of the MHSA is submitted.

10. If a student did not test and meets the following criteria, the student will be added to the enrollment at the school indicated in the student demographic file, with no test data:

    a Is actively enrolled (Active = ‘1’ in the student demographic file)

    b Is a Maine Resident (Stustatus ^= ‘4’) in the student demographic file)

    c Is a third year student (Active=’2’ in the Student demographic file)

    d Is enrolled at a ‘PUB’,‘PSP’, ‘CHA’ or ‘PSE’ school, or

    e Is enrolled at a ‘BIG’ or ‘PSN’ school and has a sending SAU]

    The student is reported as defined by the rules described in this document.

11. One common MC science item could not be included on the Braille form. Students using the science Braille form will be reported as follows:

    a The student’s response for the item will be set to ‘X’

    b The item will be excluded from calculating the student’s scaled score, achievement level, raw score, and content strand scores.

    c The content strand points earned will be set to missing for the content strand of the item. Therefore, on the ISR and Item Analysis report, the data will be blank for the affected content strand.

    d Interactive Included flag will be set to 2. The student will be included in achievement level calculations and excluded from item level and content strand aggregations.

II. Student Participation / Exclusions

A. *Test Attempt Rules*

1. A Multiple Choice (MC) item is considered attempted if

    - For Science, an A, B, C, D or a multiple response (denoted by an asterisk).

    - For Reading, Math and Writing, a ‘+’ or ‘-‘.

2. An Open Response (OR) item is considered attempted if

    a For Science, the question was not left blank.

    b For Math, (gridded responses are OR items), the question was not omitted (‘O’) or left blank.

c For writing, the writing prompt was not left blank or scored ‘00’.

d Reading does not have OR items.

3. A student attempted the test if the student attempted at least one MC item or one common OR item

B. *Not Tested Reasons (by subject)*
Please refer to the Student Demographic Specifications for how [Subject]NT is populated. If a student has more than one reason for not participating on the test, we will assign one participation code using the following hierarchy:

1. Alternate Assessment ([Subject]NT = ‘A’)*
2. Special Consideration ([Subject]NT = ‘S’)
3. Hold (SAT results on hold)
4. First Year LEP ([Subject]NT = ‘L’)
(Reading Only)
5. Did not Participate

**Students are identified as participating in the MHSA Alternate Assessment based on the MHSA Alternate Assessment Decision Rules for each subject and if they are marked as Active=’2’ in the Student Demographic file.*

C. *Student Participation Status (by subject)*

1. If the student attempted the test:

a And has a Not Tested reason of Special Consideration, Hold, or Alternate Assessment, the student will be reported with the Not Tested reason.

b Otherwise the Not Tested reason is ignored and the student will be reported as Tested on the MHSA.

2. If the student did not attempt the test:

a And has a Not Tested reason then the student will be reported with the Not Tested reason.

b Otherwise the student is reported as Did Not Participate.

D. *Student Participation Summary (by subject)*

Participation Status	**Participation Flag**	**Scaled Score**	**Achievement Level**	**ISR Report**	**ISR Text (Achievement Level)**	**Roster Code**
Alternate Assessment	C			✓*	Alternate Assessment	ALT

Special Consideration	D			✓*	Special Consideration	ASC
First Year LEP *(Reading Only)*	E			✓	First Year LEP	LEP
Hold *(Reading, Math and Writing Only)*	J			✓	SAT Results on Hold	HLD
Did not Participate	F			✓	Did not Participate	DNP
Tested MHSA without accommodations	A	✓	✓	✓	*(Earned Achievement Level)*	
Tested MHSA with accommodations (including Maine-Only)	B	✓	✓	✓	*(Earned Achievement Level)*	

** If a student has a participation status of Special Considerations and/or Alternate Assessment for all subjects assessed at the grade level, a ISR is not produced*

III. Calculations

A. *Minimum N Size*

If there are less than 10 tested participants in a subgroup (students with achievement levels), the scaled score and achievement levels are not reported. This applies to all reports with aggregations.

B. *Rounding Table*

Report	**Calculation**	**Rounded (to the nearest)**
ISR Report	Relative Achievement Level Percent	Whole value (% is displayed)
Summary of Scores	Average Scaled Score	Whole value
Summary of Student Participation	All percents	Whole value
Item Analysis Report	Multiple Choice Percent Correct Open Response Average Score	Whole value Tenth
	Content Standard Earned Average	Hundredth
	Content Standard Earned Percent	Tenth
	Content Standard Points Earned, Total Points Earned	Hundredth
Results	Percent at each achievement level, Percent of points possible, Percent of students in each Category, Scaled Score	Whole value

C. *Not Tested – Approved*

1. The Number Not Tested – Approved is the number of Not Tested Alternately Assessed, Not Tested Special Considerations, and Not Tested First Year LEP students.

D. *Not Tested – Other*

The Number Not Tested Other is the number of the Did Not Participate and Hold students

E. *Item Scores*

1. All MC items are scored using formula scoring where:

   a For science:

   - A correct response = 1 point
   - An incorrect response (or *) = -1/3 point
   - A blank response = 0 point

   b For SAT subjects:

   - '+' = 1 point
   - '-' = -1/4 point
   - 'O' or blank response= 0 point

2. The Writing Prompt is scored from 0 to 12. A blank response scores 0 point.

3. The Math Gridded Response items are scored 0 to 1.

   a '-','O' or a blank response = 0 point

   b '+' = 1 point

F. *Released Item Data*

1. The data for the released items are provided by Program Management or exist in IABS.

2. Details on how the Content Strands are derived for Science can be found later in the document under the Content Standards section.

G. Content Standards

1. The standards for Reading, Mathematics, Writing and Science are provided by program management or appear in IABS. All standards are stored in daIref.

2. Standard is displayed as Content Strand for Reading, Math and Science. Standard is displayed as SAT Item Type for Writing.

3. The SAT section is stored as 'Session' in daIref.

4. For Science, standard is calculated by concatenating Cat3 and Cat4 (i.e. the third and forth sections of ContentFramework) from IABS

5. Table of Content Standards

Subject	Repcat	Repcat Title	Standard
Reading	1	Word ID/Vocabulary	WV
	2	Literary	LT
	3	Informational	IN
Math	1	Numbers & Operations	NO
	2	Geometry & Measurement	GM
	3	Functions & Algebra	FA
	4	Data, Statistics, & Probability	DP
Writing	1	Sentence Correction	SC
	2	Usage	U
	3	Revision in Context	RC
	4	Writing Essay	ES
Science	1	D. The Physical Setting	D1,D2,D3,D4
	2	D1/D2 Space/Earth	D1,D2
	3	D3/D4 Matter/Energy/Force/Motion	D3,D4
	4	E. The Living Environment	E1,E2,E3

H. *Scaling: Assignment of Scaled Score and Achievement Level*

1. Scale Form Creation

Scaling is accomplished by defining the unique set of test forms for the grade/subject. This is accomplished as follows:

a Translate each form and position into the unique item number assigned to the form/position.

b Order the items by

- Type – multiple-choice, short-answer, constructed-response, extended-response, writing prompt.
- Form – common, then by ascending form number.
- Position

c If an item number is on a form, then set the value for that item number to ‘1’, otherwise set to ‘.’. Set the Exception field to ‘0’ to indicate this is an original test form.

d If an item number contains an ‘X’ (item is not included in scaling) then set the item number to ‘.’. Set the Exception field to ‘1’ to indicate this is not an original test form.

e Compress all of the item numbers together into one field in the order defined in step II to create the test for the student.

f Select the distinct set of tests from the student data and order them by the exception field and the descending test field.

g Check to see if the test has already been assigned a scale form by looking in the tblScaleForm table. If the test exists then assign the existing scale form. Otherwise assign the next available scale form number. All scale form numbering starts at 01 and increments by 1 up to 99.

2. Scaled Score assignment

   a For Reading, Writing and Math scaling is done using a look-up table provided by psychometrics and the SAT scaled score.

   b For Science, scaling is done using an unrounded raw score to scaled score conversion table provided by psychometrics.

   c Scaled Scores are rounded to even integers.

3. Achievement level coding:

   1 = Substantially Below Proficient

   2 = Partially Proficient

   3 = Proficient

   4 = Proficient with Distinction

IV. Report Specific Rules

A. On all reports, grade is printed as ‘High School’.

B. For achievement level data if the number of students in an achievement level does not equal 0, and the percent of students is 0 then format the percent as <1.

C. *Student Labels*

1. Student name is printed last name, first name middle initial. If a student is missing a first and last name, then report as ‘NAME NOT PROVIDED’.
2. If the student participated in the MHSA, the scaled score is printed along with the achievement level text. See section III.H.3. Otherwise, the text from section II.C is printed, based on the participation status.
3. If a student has a participation status of Special Consideration and/or Alternate Assessment for all subjects assessed at the grade level, a label is not produced (ParentLetter = ‘0’).
4. If a student is Home schooled, a label is not produced.
5. SAU code concatenated with the school code is printed at the bottom of each page of student labels.

D. *ISR Report*

1. Student name is printed first name followed by middle initial followed by last name, with spaces in between. If a student is missing a first and last name, then print ‘NAME NOT PROVIDED’.
2. If a student has a participation status of Special Considerations and/or Alternate Assessment for all subjects assessed at the grade level, an ISR is not produced (ParentLetter = ‘0’).
3. Home School students only have student and state data on this report. All school and SAU data is blanked out. School name is ‘Home Schooled’ and SAU name is the name of the student’s reported SAU.
4. If the student participated in the MHSA, the scaled score is printed along with the achievement level text. See section III.F.3. Otherwise, the text from section II.C is printed, based on the participation status.
5. All data (checkmarks, points, percents) in the content area achievement level and subcategory boxes are centered within the box.
6. The first scaled score in the display represents the lowest possible scaled score. The last scaled score in the display is the highest possible scaled score. The 3 middle scaled scores displayed represent the lowest scaled score in the 2nd, 3rd and 4th achievement levels respectively.
7. SAU code concatenated with the school code is printed at the bottom of each ISR.

E. *Item Analysis Report*

1. Group, school, SAU and state item averages will be provided for the selected group with no demographic filters applied.
2. Group and state item averages will be provided for the selected group with one or more demographic filters applied.

F. *Summary Report Package*

1. If there are less than 10 students in a school and/or SAU, only page 1 and the Summary of Student Participation pages are produced.
2. Summary of Scores

   a The Cumulative Average Scaled Score is a weighted average calculated for years where there are 10 or more tested students. If there are less than 3 years with 10 or more students, the weighted average will be left blank. The weighted average is calculated as follows:

- $$\frac{(nYear1 * ssYear1) + (nYear2 * ssYear2) + (nYear3 * ssYear3)}{nYear1 + nYear2 + nYear3}$$

- The weighted average is rounded to the nearest whole number.

3. Summary of Student Participation

   a The Current LEP category is defined as students who are identified in the student demographic file as currently receiving LEP services (LEP = ‘1’).

   b Content Area Participation

   - The numerator is the sum of students with participation statuses of Tested with accommodations, Tested without accommodations, Not Tested Alternate Assessment, and Not Tested 1st year LEP (reading only).
   - The denominator is calculated using the number enrolled minus students with a Special Consideration status.

   c Mode of Participation

   - For each Mode of Participation group (Participated with Accommodations, without Accommodations, Participation through PAAP, and the non-Participation groups), the percents are calculated using a denominator of the total number of students enrolled. The sum of the N’s for these 6 groups is equal to the number total number of students enrolled.
   - For each subgroup within the modes (Identified disability and LEP), the percents are calculated using a denominator that is the number of students in that particular group (Mode). LEP is the number of Current LEP students (LEP = 1)

4. Student Questionnaire

   a Only tested students will be included in the calculations.

   b Percent of students in this category is computed by the number of tested students that selected that response/number of tested students with a single response for the question *100. Students are considered to have a single response, if their response is not blank or ‘*’.

5. Results

a The Cumulative Total N for an achievement level on the results page is calculated as follows:

- $nYear1_i + nYear2_i + nYear3_i$

  with i representing the achievement level

b The Cumulative Total Percent for an achievement level is calculated as follows:

- $100*\left(\frac{CumulativeTotalN}{\sum_1^4 CumulativeTotalN_i}\right)$

  with i representing each achievement level

c If there are less than 3 years with 10 or more students, the Cumulative Total N and Percent for each achievement level will be left blank.

d The Cumulative Total N and Percent is not calculated for Science this year because there are not 3 years of data for that content area.

e The Current LEP Yes category is defined as students who are identified in the student demographic file as currently receiving LEP services. (LEP = ‘1’). Current LEP No category is defined as students who are identified in the student demographic file as Not Currently receiving LEP services (LEP not = ‘1’)

Shipping Information

A. *School Products (ReportFor = 1)*

1. The ISR reports will be class-packed at the printer. Each pack will contain 1 set of ISR reports for that school.
2. The student results labels will be class-packed at the printer. Each pack will contain 1 set of student labels for that school.

Report Description	Grade	Report Type	Content Code	Subject	Quantity
Parent Report	11	02	00	Math, Reading, Writing, Science	1
Student Results Label	11	03	00	Math, Reading, Writing, Science	1

V. Data Requirements Interactive Reporting

A. Student Level

1. Students will be loaded into the Interactive System based on the Interactive flag in tblStuDemo. Students with Interactive flag set to 0 will not be loaded into the system. Students with Interactive set to 1 will be loaded.

a Students with StuStatus value of 1, 2, 3 will have the Interactive flag set to 0.
b Grade 11 students who are not marked as Active='2' in the Student Demographic file will have the interactive flag set to 0.
c All others will have Interactive=1.

2. The Included flag will determine which students are included in school level aggregations. Students with Included=0 are excluded from all school level aggregations. Students with Included=2 will be included in Performance Level aggregations and excluded from raw score aggregations (item, subcategory, and total raw score). Students with Included=1 will be included in all school level aggregations.

a Included = 1:

- The student has a Participation Status of A or B and
- The student took the released form of the SAT
- The student is included in school level aggregations

b Included = 2:

- The student has a Participation Status of A or B and took a non released form of the SAT or
- The student has a Participation Status of A or B and is identified as Braille.
- The student is included in school level aggregations
- Data Analysis will blank out all SAT items and Points Earned so they will not be displayed on the Item Analysis Report. The student will receive a Scaled Score and Achievement Level.
- IS will print 'NRF' in the Total Points Earned column, which means Not Released Form.

c Included = 0:

- The student did not participate in the MHSA (Participation status is not A or B).
- The student is excluded from school level aggregations.
- Data Analysis will blank out all items, Points Earned and the Scaled Score so they will not be displayed on the Item Analysis Report for students who did not participate in the MHSA (Participation status is not A or B).
- IS will print the 3 character not tested code in the Achievement Level column.
-

B. Item Average Calculations:

a Students are included in item average calculations if Included = 1

C. Achievement Level Summaries:

Students are included in achievement level summaries if Included = 1 or 2

1. Longitudinal Data
   - a Only Interactive flag=1 students will be loaded.
   - b The complete achievement level name or not tested reason will be stored.
   - c Results will be loaded for MHSA0809 and MHSA0910.

D. Aggregate Level

1. Data Analysis will compute Item Averages for the whole group only at the School and SAU Levels.
2. Data Analysis will compute Item Averages for all of the filter combinations that exist at the State Level.
3. Data Analysis will create a lookup table with all of the possible filter combinations. It will contain the variable Filter with length 5. Each position represents one of the filter variables. It will contain all the possible combinations of the values plus nulls for when variables are not selected. The first position will be Gender, second Ethnic, third IEP, fourth LEP, and fifth EconDis.
4. Data Analysis will compute Item Averages, Achievement Level Summary, and Item Summary data for the filter combinations for a sample of schools for quality assurance review.
   - a For this sample, percents will be rounded to the nearest whole number and open response average scores will be rounded to the nearest tenth.
   - b For the Item Summary data, item responses other than A, B, C, and D will be counted in the IR column.

VI. Data File Rules

A. Refer to file layouts for data elements and structure.

B. Grade 11 students who are not marked as Active='2' in the Student Demographic file are not included in any post reporting aggregation files.

C. *State Student Raw Data Files*

1. Exclude students with StuStatus=1, 2, 3, or 4
2. Only students from 'PUB', 'PSP', 'CHA' and 'PSE' schools are included, or if they have a sending SAU.
3. Students with all participation statuses are included.
4. There are two files per grade; one with names and one without.
5. Field test item responses are not displayed.
6. Science items are not included in the Without science files.

7. Data are ordered by Student Grade, SAU code, School code, last name, and first name.

D. *State Student Scored Data Files*

1. Exclude students with StuStatus=1, 2, 3, or 4
2. Only students from ‘PUB’, ‘PSP’, ‘CHA’ and ‘PSE’ schools are included, or if they have a sending SAU.
3. Students with all participation statuses are included.
4. There are two files per grade; one with names and one without.
5. Field test item responses are not displayed.
6. The files are ordered by Student grade, SAU code, School code, last name, and first name.

E. *State Accountability Student Results Data*

Specifications are in HS Accountability Decision Rules

F. *School/SAU Student Results Files*

1. Exclude students with StuStatus=1, 2 or 3
2. Only ‘PUB’, ‘PSP’, ‘CHA’ and ‘PSE’ school SAUs will receive SAU files.
3. Students with all participation statuses are included.
4. A student with a sending SAU is included in both the tested SAU file and the sending SAU file.

G. *Press Release*

1. The data reported in these files are the number of students tested, the number and percent of students performing at each achievement level, and the average scaled score.
2. The SAU file is only produced for ‘‘PUB’, ‘PSP’, ‘CHA’ and ‘PSE’ SAUs and include students with participation statuses of Tested with or without accommodations. A student with a sending SAU is aggregated only to the sending SAU.
3. The school file is only produced for ‘PUB’, ‘BIG, ‘PSP’, ‘CHA’ and ‘PSE’ schools, and include students with participation statuses of Tested with or without accommodations.
4. Schools or SAUs that have < 10 included students will only include data for the number of students tested.

H. *State Accommodation Frequency Report*

1. The data reported in these files are the counts of each SAT accommodation and the Maine-Only accommodation.
2. Exclude students with stustatus=1, 2, 3, or 4 and students with not tested participation status

I. *Top 50 HS Students*

Each student is rank ordered by scaled score in all subjects. These rankings are averaged for each student who participated in all subjects. The students are then ordered by the average ranking and the top 50 students are identified.

J. *State Standard Deviations & Average Scaled Scores*

1. Exclude students with StuStatus=1, 2, 3, or 4
2. This file includes students from 'PUB', 'PSP', 'CHA' and 'PSE' schools, or if they have a sending SAU.
3. Students with participation statuses of Tested with or without accommodations are included.
4. The data reported in these files are the number of students tested, the average scaled score, and the standard deviations for the following subgroups:

   a Identified Disability, No Identified Disability

   b LEP (Currently receiving LEP services), Not LEP

   c Economically Disadvantaged, Not Economically Disadvantaged

   d Migrant, Not Migrant

   e Gender

   f Ethnicity

   g Title 1, Not Title 1

   h Total (All students)

K. *Minimally Statistically Significant Differences for Scaled Scores*

The data reported in this file are the number of scaled score points denoting minimally statistically significant differences for average school/SAU results. This is calculated by psychometrics.

L. *Standard Error of Measurement (Reliability)*

1. This file consist of three worksheets
2. Each worksheet contains the number of students tested, the number of possible raw score points, the minimum, maximum, mean, standard deviation, reliability, and the SEM of the raw score.
3. The subgroup categories are same as the ones reported on the “Results By Reporting Subgroups” page of the “Summary Report Package” report.
4. Only Science is included.
5. Only students included in State level aggregations are included.

M. Score Ranges (Science Only)

Contain two worksheets: *MHSAYYYYScaledScoreRanges* and *MHSAYYYYRawScoreRanges*

N. *State Student Questionnaire*

1. The data reported in this file are the responses to the Student questionnaire, performance levels and scaled scores
2. Exclude students with stustatus=1, 2, 3, or 4
3. Only students who receive a performance level in at least one subject are included

O. Department Chair and Principal Questionnaire Raw Data

1. One CSV file will be created containing raw Department Chair Questionnaire data.
2. One CSV file will be created containing raw Principal Questionnaire data.

P. Department Chair/Principal Questionnaire Frequency Distribution

One CSV file will be created containing the distribution of responses of Department Chair/Principal Questionnaire raw data.

VII. Data File Table

(YYYY indicates year, SSS indicates subject)

File	Naming Convention
State Student Raw Data	MHSA*YYYY*StateStudentRawDataHS.csv MHSA*YYYY*StateStudentRawDataNoNamesHS.csv MHSA*YYYY*StateStudentRawDataLayout.xls MHSA*YYYY*StateStudentRawDataHS_NoSci.csv MHSA*YYYY*StateStudentRawDataNoNamesHS_NoSci.csv MHSA*YYYY*StateStudentRawDataLayout_NoSci.xls
State Student Scored Data	MHSA*YYYY*StateStudentScoredDataHS.csv MHSA*YYYY*StateStudentScoredDataNoNamesHS.csv MHSA*YYYY*StateStudentScoredDataLayout.xls
School/SAU Student Results	SAU:MHSA*YYYY*StudentHS*SSS_[SAUcode]*.csv Sch:MHSA*YYYY*StudentHS*SSS_[SAUcode+schcode]*.csv MHSA*YYYY*StudentReleasedItemLayout.xls
Press Release	MHSA*YYYY*SchoolPressReleaseHS.csv MHSA*YYYY*DistrictPressReleaseHS.csv MHSA*YYYY*PressReleaseLayout.xls
State Accommodation Frequency Report	MHSA*YYYY*AccommodationHS.xls
Top 50 HS Students	MHSA*YYYY*Top50.rtf
Standard Deviations & Average Scaled Scores for MHSA subgroups	MHSA*YYYY*StandardDeviationHS.xls
Standard Error of Measurement	MHSA*YYYY*Reliabilty.xls
Minimally Significant Differences for Scaled Scores	MHSA*YYYY*SignificantDifferenceChart.xls
Score Ranges (Science Only)	*MHSAYYYYScoreRanges.xls*
Scaled Score Lookup (Science Only)	MHSA*YYYY*ScaledScoreLookup.xls
MHSA School Mailing List	MHSA*YYYY*SchDisList.xls
State Student Questionnaire	MHSA*YYYY*StateStudentQuestionnaire.csv MHSA*YYYY*StateStudentQuestionnaireLayout.xls
Department Chair and Principal Questionnaire Raw Data	MHSA*YYYY*DepartmentChairQuestionnaireRaw.csv MHSA*YYYY*DepartmentChairQuestionnaireRawLayout.xls MHSA*YYYY*PrincipalQuestionnaireRaw.csv MHSA*YYYY*PrincipalChairQuestionnaireRawLayout.xls
Department Chair/Principal Questionnaire Frequency Distribution	MHSA*YYYY*DepartmentChair_PrincipalQuestionnaireFreqLayout.xls MHSA*YYYY*DepartmentChair_PrincipalQuestionnaireFreq.csv
State Accountability Student Results Data	Specifications are in HS Accountability Decision Rules